グリーンなスパコンはエクサスケールの夢を見るか
- TSUBAME2.0を例にして -

東京工業大学 学術国際情報センター
教授
松岡　聡

PC Cluster Consortium招待講演
2010年12月10日

TSUBAMEの歴史

2006
TSUBAME 1.0
85TFlops/1.1PB

アジアNo.1
「みんなのスパコン」
x86 CPU+アクセラレータ

2007
TSUBAME 1.1
100TFlops/1.6PB

ストレージ・アクセラレータ
増強

2008
TSUBAME 1.2
160TFlops/1.6PB

GPUアクセラレータ
680枚増強

2010/11/01
TSUBAME 2.0
2.4PFlops/7.1PB
日本初のペタコン、TSUBAME1.0の30倍の性能

- TSUBAME初の完全リプレース

今後のペタ級マシン

Inst/Agency/Country(	Name	Machine	Peak Perf
ORNL/DoE/US	Jaguar Upgrade	Cray XT5	2.3PF
Tennessee大学/NSF/US 2009	Cracken	Cray XT5	1PF
Julich/欧州(ドイツ)	Jugene	IBM BG/P	1PF
中国・防衛大学	天河 (Tihanhe 1)	GPU Cluster/Dawning	1.2PF
中国・深圳国立スパコン	星雲 (Nebulae)	GPU Cluster/???	3PF
日本・東工大	TSUBAME2.0	GPU Cluster/HP-NEC	2.4 PF
LBNL/DoE/US 2010	Hopper	Cray XE6	1.3PF
中国・防衛大学	天河 (Tihanhe 1-A)	GPU Cluster/Dawning	5 PF
欧州PRACE計画・仏CEA	Tera 100	Nehalem-EX Cluster/Bull	1.25PF
ORNL/DoE/US	Jaguar Upgrade 2	Cray XE6 +GPU	20PF
NCSA/NSF/US	Blue Waters	IBM Power7 server	10PF
LLNL/DoE/US	Sequoia	IBM BG/Q	20PF
ArgonneNL/DoE/US 2011-12	???	IBM BG/Q	10PF
日本・理研	「京」	富士通 Venus 専用設計	10PF
日本・筑波大	HA-PACS	GPU Cluster/HP-NEC	1PF
欧州ペタコン群/PRACE計画	???	IBM, Cray等	~PF x 4~5
中国	4~6個所	???Dawning?	合算数十PF以上

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

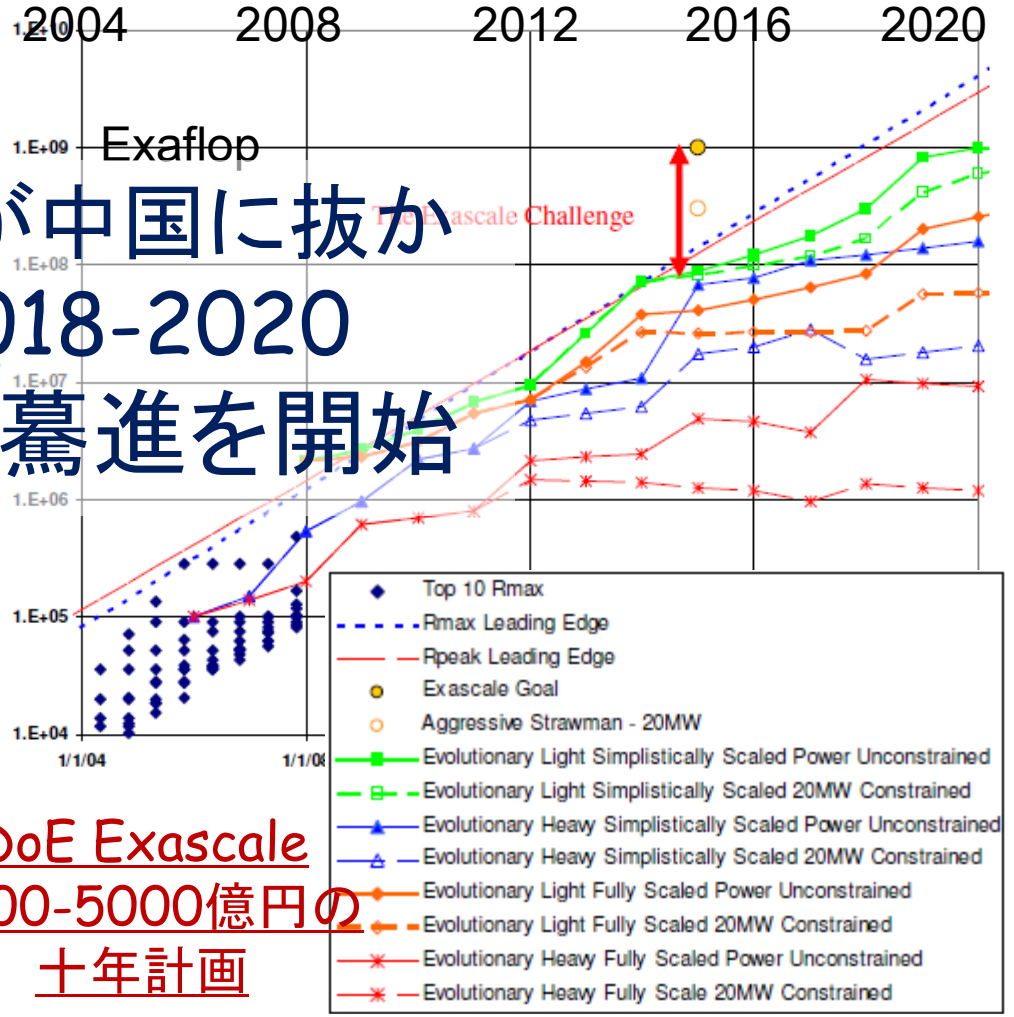
Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod

Petaを達成したが中国に抜かれた米国は2018-2020 Exa(10^{18})flopへ驀進を開始

Peter Koggeらによる300ページのDoD Exascaleシステムのレポート



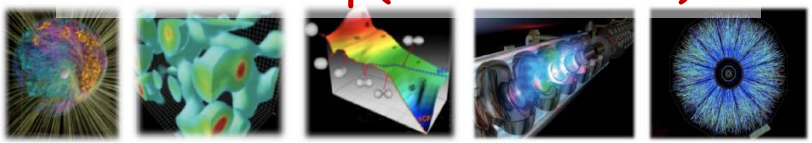
DoE Exascale 2000-5000億円の十年計画

Exa-scale Computational Resources

(slide courtesy Martin Savage)

- Meeting structured around present DOE Office of Science areas of effort

6アプリ分野のExascale Workshop(2008-2009)

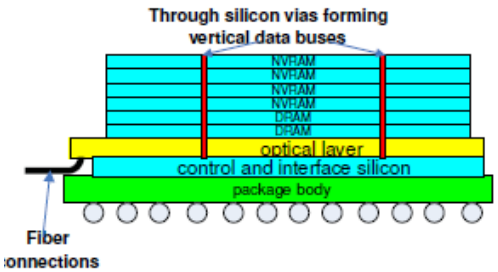


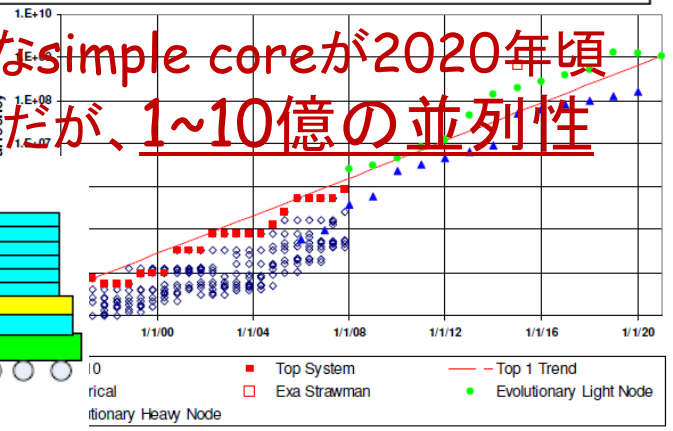
Nuclear Astrophysics | Cold QCD and Nuclear Forces | Nuclear Structure and Reactions | Accelerator Physics | Hot and Dense QCD

- Exa-scale computing is REQUIRED to accomplish the Nuclear Physics mission in each area
- Staging to Exa-flops is crucial :
  - 1 Pflop-yr to 10 Pflop-yrs to 100 Pflop-yrs to 1 Exa-flop-yr (sustained)

Paul Messina June 28, 2009

軽量なsimple coreが2020年頃有望だが、1~10億の並列性





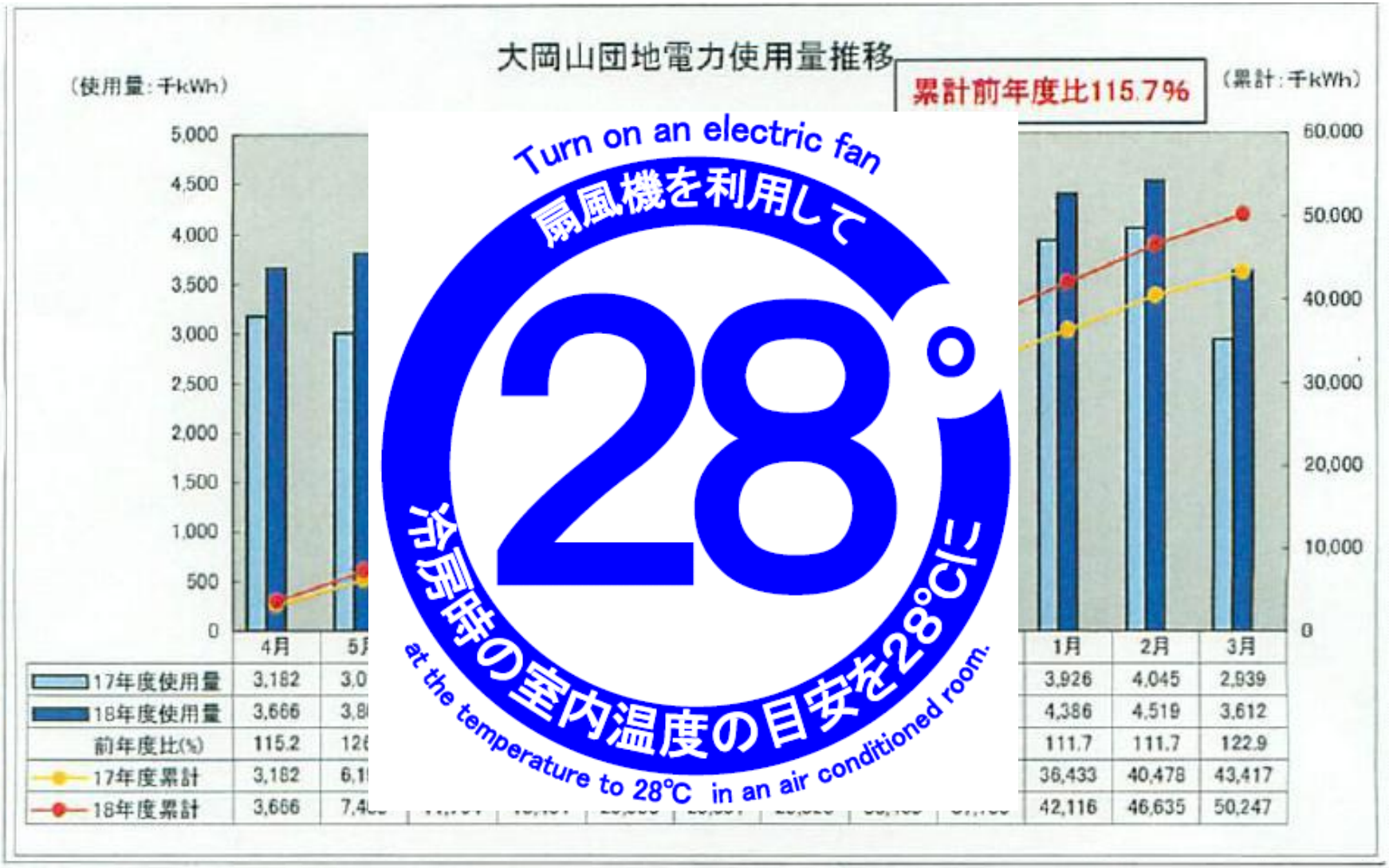
US DoE Exascale Roadmap

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

ペタ～エクサへのスケーリングのロードブロック

- 「10億並列へ」は勇ましいが。。。
  - 電力・エネルギー
  - (強)スケーリングの欠落
  - N^2 vs. N 問題により深まるメモリ階層 (I/O含む)
  - 極端に低まる信頼性と実行不能性
  - プログラミングや実行モデル

ある日施設課に呼び出され。。。



	4月	5月	1月	2月	3月
17年度使用量	3,182	3,0	3,926	4,045	2,939
18年度使用量	3,666	3,8	4,386	4,519	3,612
前年度比(%)	115.2	12	111.7	111.7	122.9
17年度累計	3,182	6,1	36,433	40,478	43,417
18年度累計	3,666	7,4	42,116	46,635	50,247

Biggest Problem is Power…

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS/ Watt	Watts/ CPU Core	Ratio c.f. TSUBAME
TSUBAME(Opteron)	10480	800,000	50,400	63.00	76.34	
TSUBAME2006 (w/360CSs)	11,200	810,000	79,430	98.06	72.32	
TSUBAME2007 (w/648CSs)	11,776	820,000	102,200	124.63	69.63	1.00
Earth Simulator	5120	6,000,000	40,000	6.67	1171.88	0.05
ASCI Purple (LLNL)	12240	6,000,000	77,824	12.97	490.20	0.10
AIST Supercluster (Opteron)	3188	522,240	14400	27.57	163.81	0.22
LLNL BG/L (rack)	2048	25,000	5734.4	229.38	12.21	1.84
Next Gen BG/P (rack)	4096	30,000	16384	546.13	7.32	4.38
TSUBAME 2.0 (2010Q3/4)	160,000	810,000	1,024,000	1264.20	5.06	10.14

TSUBAME 2.0 x24 improvement in 4.5 years…? ➔ ~ x1000 over 10 years

2016年*ULP-HPC*技術により
デスクサイド・ペタスケールコンピューティングへ
JST-CREST Ultra Low Power HPCにおける基礎研究

2006年東工大TSUBAME
アジア一位：85TeraFlops,
> 10,000 CPU, 1.2 MegaWatt, 350m^2

ペタコン技術の1000倍のダウンスケール：
どうやって？

2016年デスクサイドワークステーション
~100TeraFlops, 1.5KiloWatt,
350cm^2

単純なスケーリングでは達成不可

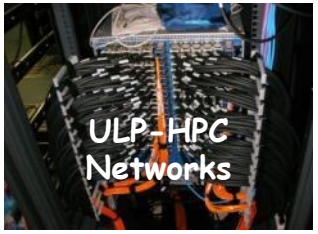
クロック上昇不可
「マルチコア」でもせいぜい数十~100倍

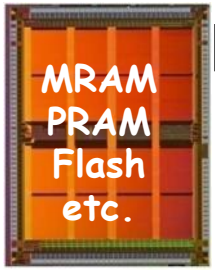
ULP-HPC
1000倍(あとx10)の技術ブレークスルーを行う情報科学研究

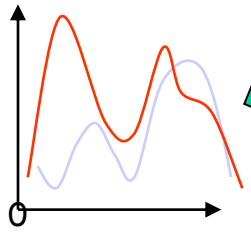
「ペタスケールシミュレーションがエンジニアにとって日常的に」

JST CREST *ULP-HPC*全体スキーム

ULP-HPC
SIMD-Vector
(GPGPU, etc.)







省電力高性能
ソフト機構・モデル化

(提案3) 自動チューニング共通基盤
(須田@東大)

モデルと実測の Bayes 的融合

- Bayes モデルと事前分布

$$y_i \sim N(\mu_i, \sigma_i^2)$$

モデルによる所要時間の推定

$$\mu_i \mid \beta, \sigma_i^2 \sim N(x_i^T \beta, \sigma_i^2 / \kappa_0)$$

$$\sigma_i^2 \sim \text{Inv-}\chi^2(v_0, \sigma_0^2)$$

- n 回実測後の事後予測分布

所要時間の実測データ

$$y_i \mid (y_{i1}, y_{i2}, \cdots, y_{in}) \sim t_{v_n}(\mu_{in}, \sigma_{in}^2 \kappa_{n+1} / \kappa_n)$$

$$v_n = v_0 + n, \quad \kappa_n = \kappa_0 + n, \quad \mu_n = (\kappa_0 x_i^T \beta + n\bar{y}_i) / \kappa_n$$

$$v_n \sigma_n^2 = v_0 \sigma_0^2 + \sum (y_m - \bar{y}_i)^2 + \kappa_0 n (\bar{y}_i - x_i^T \beta)^2 / \kappa_n$$

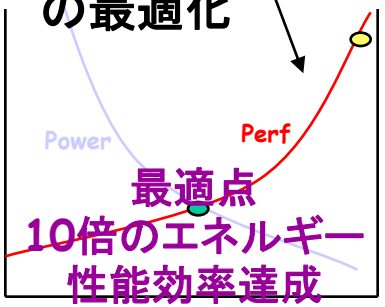
$$\bar{y} = \frac{1}{n} \sum y_i$$

ABCLibScript: アルゴリズム選択

```
!ABCLib$ static select region start
!ABCLib$ parameter (in CacheS, in NB, in NPrc)
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$    (2.0d0*CacheS*NB)/(3.0d0*NPrc)
            対象1（アルゴリズム1）
!ABCLib$    select sub region end
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$    (4.0d0*CacheS*dlog(NB))/(2.0d0*NPrc)
            対象2（アルゴリズム2）
!ABCLib$    select sub region end
!ABCLib$ static select region end
```

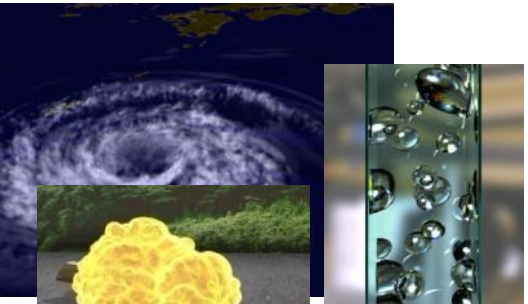
実行起動前自動チューニング指定、アルゴリズム選択処理の指定
コスト定義関数で使われる入力変数
コスト定義関数
対象領域1、2

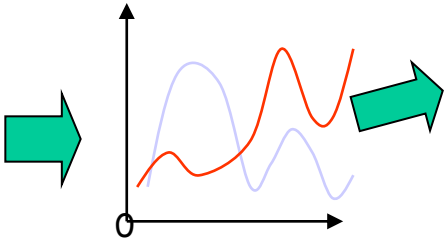
電力性能比の最適化



最適点
10倍のエネルギー性能効率達成

(提案1) ULP-HPC用次世代SW/HW要素利用技術およびモデル化
(松岡@東工大・本多@電通大、鯉渕@NII)





省電力高性能
アルゴリズム・モデル化

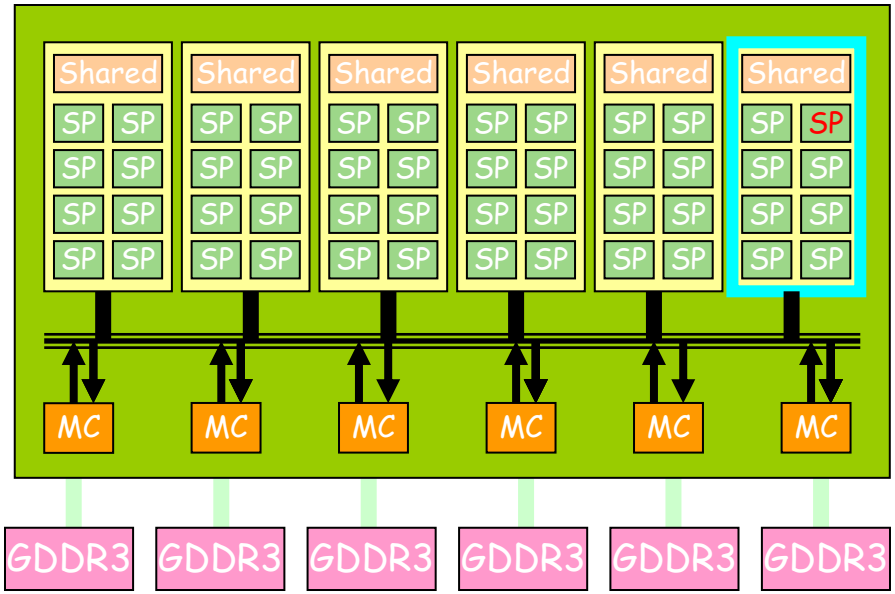
2016年 TSUBAME
1/1000に

(提案2) 超省電力型のHPCアプリケーション及びアルゴリズム(青木@東工大)

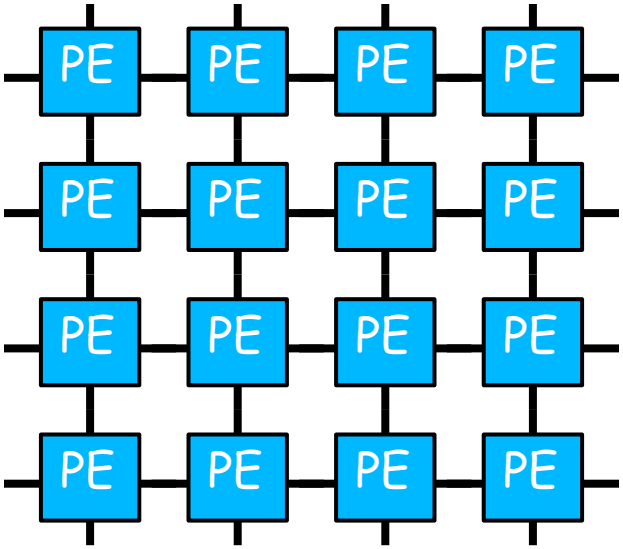
省電力化の手法と有効性

情報基盤/手法	エンタープライズ・ビジネス・クラウド	HPC
仮想化による統合化 (Server Consolidation)	○	×
DVFS (Differential Voltage/Frequency Scaling)	○	△
新デバイス	△ (コストや継続性)	○
新アーキテクチャ	△ (コストや継続性)	○
冷却技術	○	△ (ただし高熱密度)

GPU（Multithreaded Vector）vs. Standard Many Cores?



vs.



- 今後スパコンが巨大化するにつれ、強スケーリング（*strong scaling*）が選択肢において重要な決定要素に

DOE のキーアプリケーション群

(following slides courtesy John Shalf @ LBL NERSC)

NAME	Discipline	Problem/Method	Structure
MADCAP	Cosmology	CMB Analysis	Dense Matrix
FVCAM	Climate Modeling	AGCM	3D Grid
CACTUS	Astrophysics	General Relativity	3D Grid
LBMHD	Plasma Physics	MHD	2D/3D Lattice
GTC	Magnetic Fusion	Vlasov-Poisson	Particle in Cell
PARATEC	Material Science	DFT	Fourier/Grid
SuperLU	Multi-Discipline	LU Factorization	Sparse Matrix
PMEMD	Life Sciences	Molecular Dynamics	Particle

アプリケーションにはバンド幅？レーテンシ？
Latency Bound vs. Bandwidth Bound?

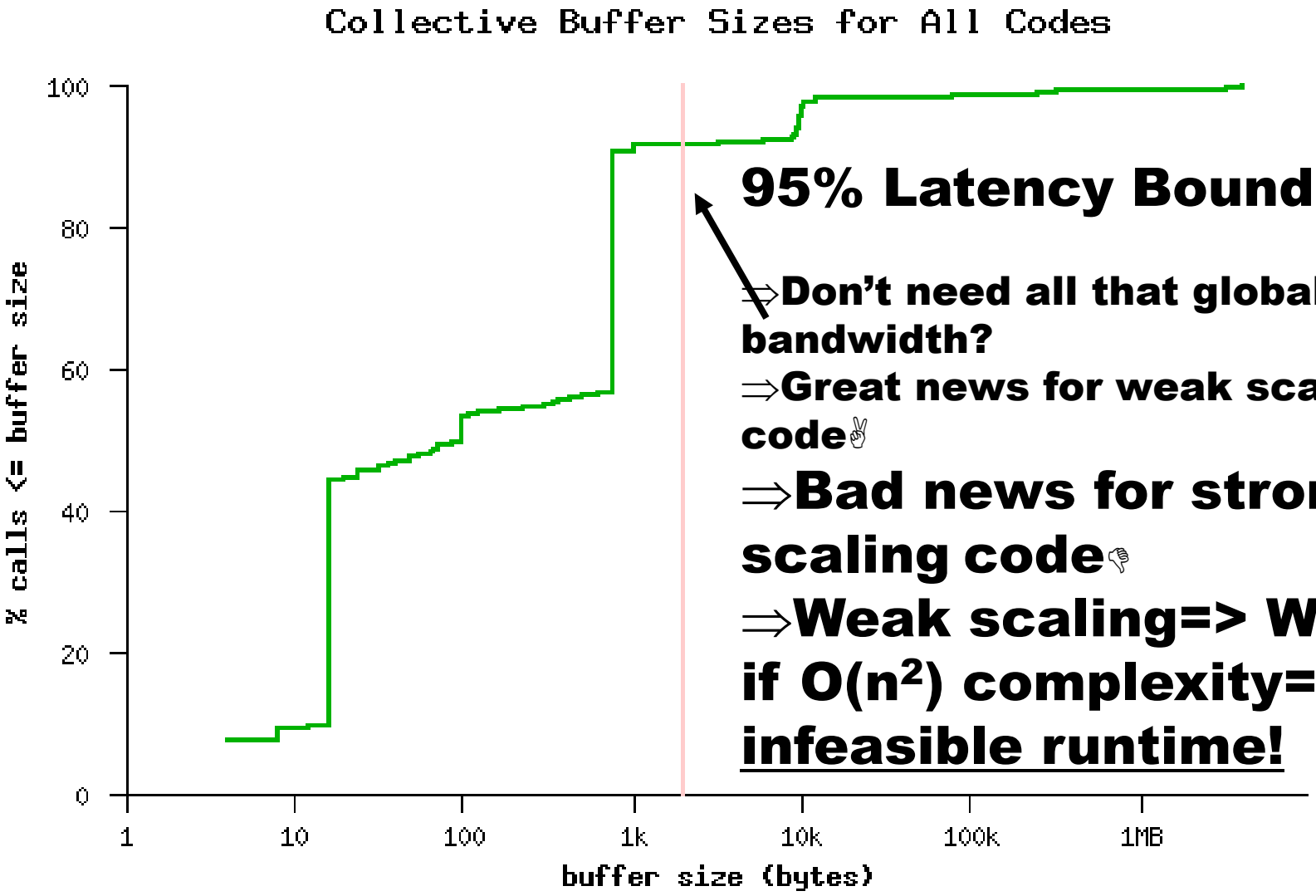
- How large does a message have to be in order to saturate a dedicated circuit on the interconnect?
  - $N^{1/2}$ from the early days of vector computing
  - Bandwidth Delay Product in TCP

System	Technology	MPI Latency	Peak Bandwidth	Bandwidth Delay Product
SGI Altix	Numalink-4	1.1us	1.9GB/s	2KB
Cray X1	Cray Custom	7.3us	6.3GB/s	46KB
NEC ES	NEC Custom	5.6us	1.5GB/s	8.4KB
Myrinet Cluster	Myrinet 2000	5.7us	500MB/s	2.8KB
Cray XD1	RapidArray/IB4x	1.7us	2GB/s	3.4KB

- Bandwidth Bound if msg size > Bandwidth*Delay
- Latency Bound if msg size < Bandwidth*Delay
  - Except if pipelined *(unlikely with MPI due to overhead)*
- W/HW DMA a few 100ns but not much more

(Original slide courtesy John Shalf @ LBL)

多くの実問題は実はレーテンシバウンド
–小規模メッセージパッシングプロセッサの問題–



95% Latency Bound!!!

⇒Don't need all that global NW bandwidth?
⇒Great news for weak scaling code✌
⇒Bad news for strong scaling code👎
⇒Weak scaling=> What if $O(n^2)$ complexity=> infeasible runtime!

(Original slide courtesy John Shalf @ LBL)

ペタからエクサへのスケーリング
強スケーリング達成のためには

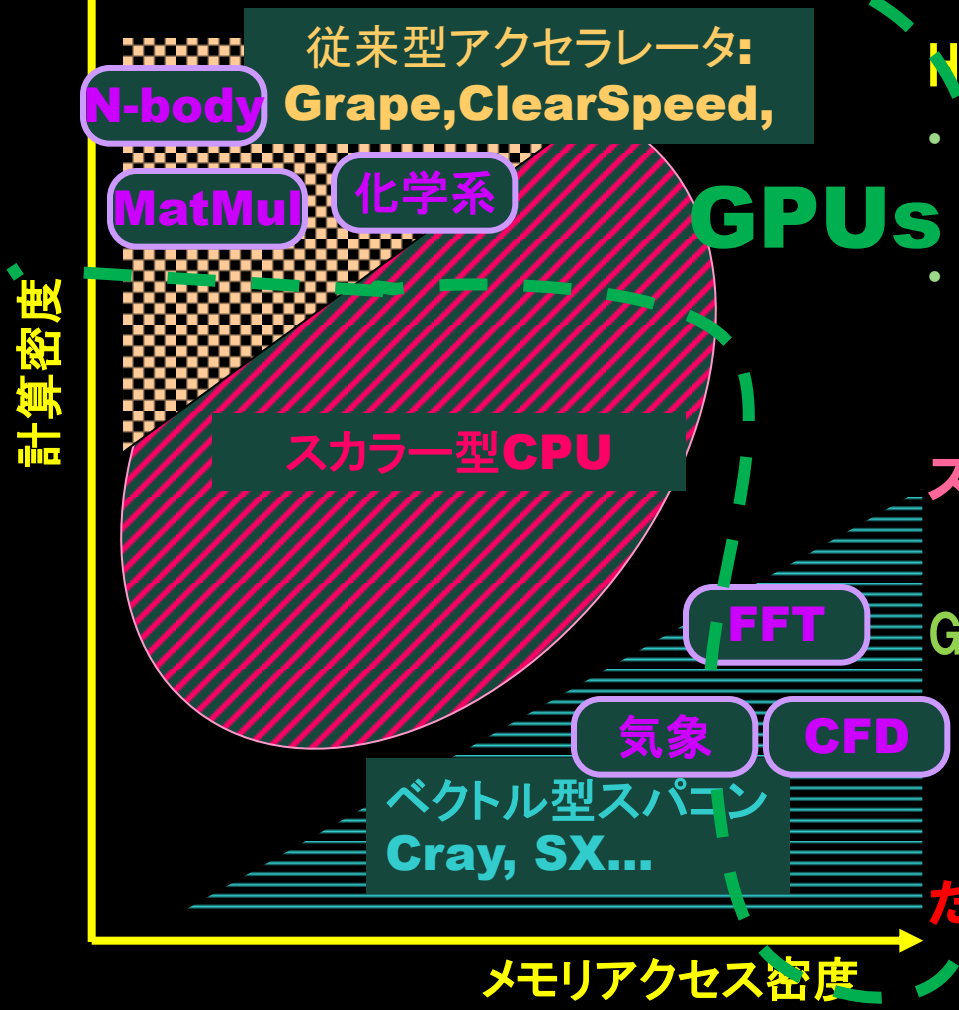
- ***レーテンシをなるべく短く***
  - Extreme multi-core incl. vectors
  - "Fat" nodes, exploit short-distance interconnection
  - Direct cross-node DMA (e.g., put/get for PGAS)
- ***又はレーテンシ隠し(高バンド幅＋大量のスレッド)***
  - Dynamic multithreading (Old: dataflow, New: GPUs)
  - Trade Bandwidth for Latency (so we do need BW…)
  - Departure from simple mesh system scaling
- ***レーテンシに敏感なアルゴリズムを変更***
  - From implicit Methods to direct/hybrid methods
  - Structural locality, extraploation, stochastics (MC)
  - Still may require global bandwidth for implicit solvers

GPUs as Commodity Massively Parallel Vector Processors

- E.g., NVIDIA Tesla, AMD Firestream
  - High Peak Performance > 1TFlops
    - Good for tightly coupled code e.g. Nbody
  - **High Memory bandwidth (>100GB/s)**
    - **Good for sparse codes e.g. CFD**
  - **Low latency over shared memory**
    - **Thousands threads hide latency w/zero overhead**
  - **Slow and Parallel and Efficient vector engines for HPC**
  - Restrictions: Limited non-stream memory access, PCI-express overhead, programming model etc.

How do we exploit them given vector computing experiences?

新世代のベクトル計算機としてのGPU



NPCのワークロードとして、二つのタイプ

- 計算密度が高い「密問題」
  →従来型のアクセラレータが得意
- メモリアクセス密度が高い「粗問題」
  →ベクトル型スパコンが得意

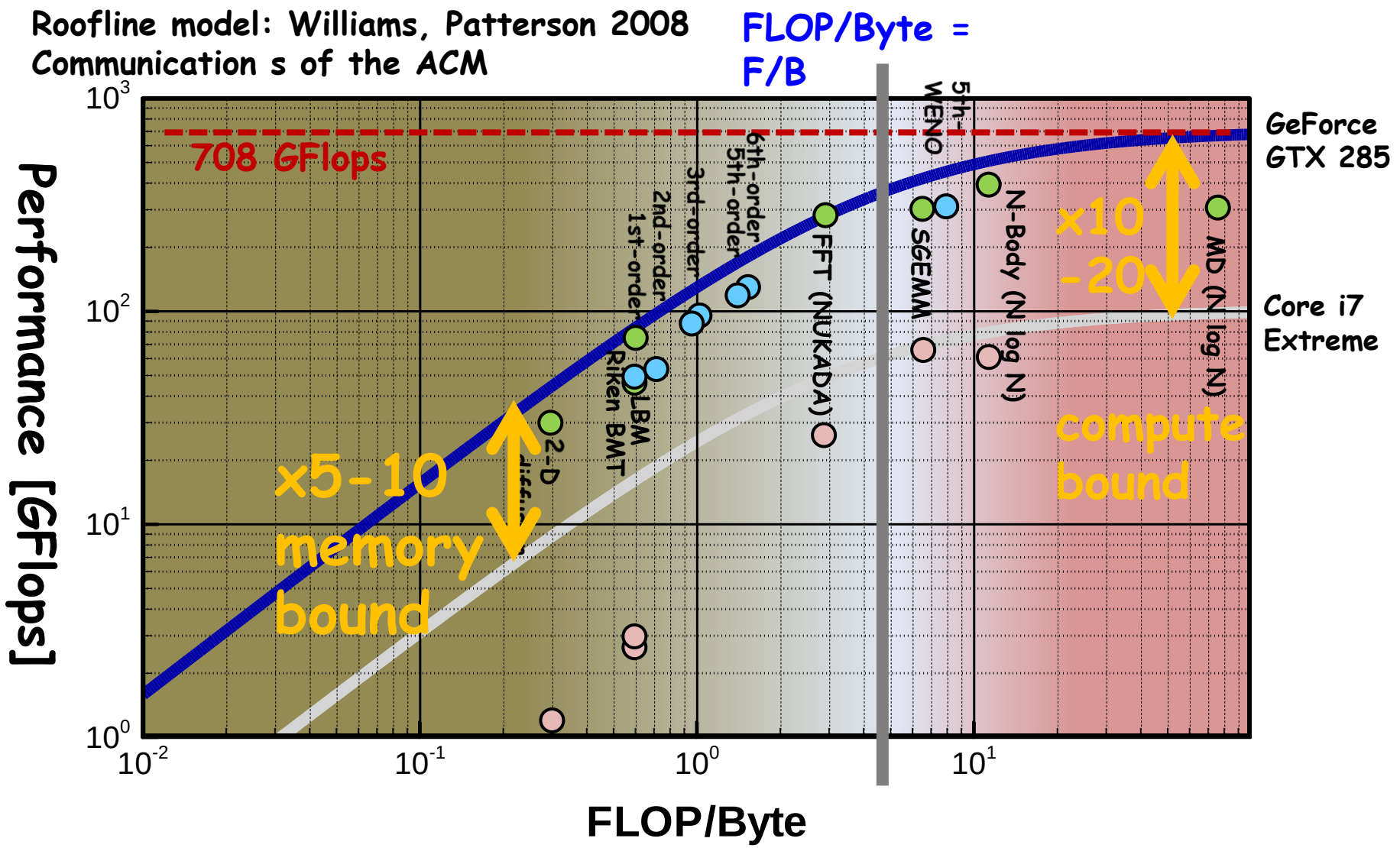
スカラー型CPUはどちらもそこそこ
→高性能を得るために巨大並列化

GPUは、新世代のベクトルプロセッサと、計算密度が高いアクセラレータとしての両面を持つ
→効率の良いスパコンの主要素

ただし、少ないメモリ量・CPUや他GPUとの通信・GPU向アルゴリズムやアプリ・ベクトル並列のプログラミング手法・システムソフトウェア等の技術課題

東工大GSICでの研究開発

GPU vs. CPU Performance



TSUBAME 1.2への進化 (GPU等拡張) 2008年10月

GSIC
Global Scientific Information and Computing Center

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

NEC SX-8i

Storage
1.5 Petabyte (Sun x4500 x 60)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s aggregate I/O BW

10Gbps+External NW

Unified Infiniband network

500GB 48disks

10,000 CPU Cores
300,000 SIMD Cores
~900TFlops-SFP,
~170TFlops-DFP
80TB/s Mem BW (x2 ES)

Sun x4600 (16 Opteron Cores)
32~128 GBytes/Node
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

GCOE TSUBASA
Harpertown-Xeon
90Node 720CPU
8.2TeraFlops

NEW: co-TSUBAME
72Node 586CPU (Low Power)
~5TeraFlops

PCI-e

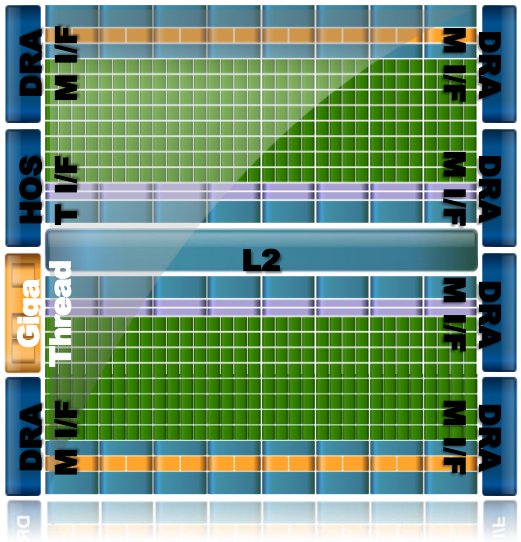
ClearSpeed CSX600
SIMD accelerator
360 648 boards,
35
52.2TeraFlops

NVIDIA

Nvidia Tesla S1070: 170台, 総計 680カード
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)


680 Unit Tesla Installation…
While TSUBAME in Production Service (!)

TSUBAME2.0のノードアーキテクチャ
GPU+CPUによるスカラー・ベクトル混合型アーキテクチャ



NVIDIA Tesla M2050 GPU
515GFlops/ 448 CUDA Cores
3GBメモリ容量、**150GB/s**メモリバンド幅
新世代のベクトルプロセッサ

GPU
Vector

GDDR5
150GB/s

Core
Dataset
Vector-
Parallel

CPU Scalar
- OS
- Services
- Legacy
- 非均質疎行列系

PCIe 2 x16
8GB/s x n

IOH

IB QDR
HCA

40Gbps IB x 2

Flash 200GB
>450MB/s

ネットワーク・I/O強化のためのカスタムMB設計

Highlights of TSUBAME 2.0 Design (Oct. 2010) w/NEC-HP

- 2.4 PF Next gen multi-core x86 + next gen GPGPU
  - 1432 nodes, Intel Westmere/Nehalem EX
  - 4224 NVIDIA Tesla (Fermi) M2050 GPUs
  - ~100,000 total CPU and GPU "cores", High Bandwidth
  - **1.9 million "CUDA cores", 32K x 4K = 130 million CUDA threads(!)**

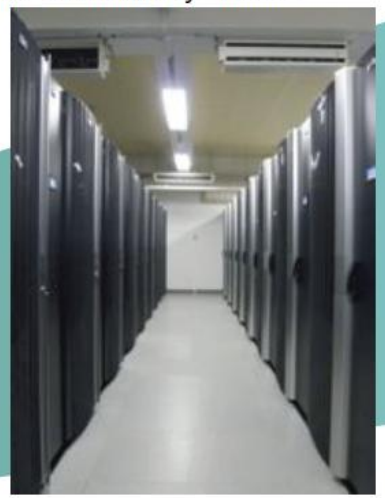
- 0.72 Petabyte/s aggregate mem BW,
  - Effective 0.3-0.5 Bytes/Flop, restrained memory capacity (100TB)
- Optical Dual-Rail IB-QDR BW, full bisection BW(Fat Tree)
  - 200Tbits/s, Likely fastest in the world, still scalable
- Flash/node, ~200TB (1PB in future), 660GB/s I/O BW
  - >7 PB IB attached HDDs, 15PB Total HFS incl. LTO tape
- Low power & efficient cooling, comparable to TSUBAME 1.0 (~1MW); PUE = 1.28 (60% better c.f. TSUBAME1)
- Virtualization and Dynamic Provisioning of Windows HPC + Linux, job migration, etc.

TSUBAME2.0 2010年11月1日稼働開始

TSUBAME2.0: A GPU-centric Green 2.4 Petaflops Supercomputer

Tsubame 2.0: "Tiny" footprint, very power efficient
- Floorspace less than 200m² (2,100 ft²)
- Top-class power efficient machine on the Green 500

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes



2.4 PFLOPS
80 TB

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB

Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS

GPUs(Tesla M2050)
515 GFLOPS
3 GB

Integrated by NEC Corporation

TSUBAME2.0の特徴

1. 世界一クラスのペタコン: 倍精度2.4ペタフロップス

- 最新型GPU・CPUによるベクトル・スカラー混合アーキテクチャ:高い計算性能とバンド幅
  - 2.4 Petaflops, メモリバンド幅0.72ペタバイト/s (地球シミュレータの4.3倍)
- 世界最高速クラスの200テラビット級のバイセクションバンド幅を実現した光ネットワーク
- 最新のSSDなどを活用した多階層ストレージによる15PBの大規模化と高速化 (0.66TB/s)

2. 世界一環境「グリーンスパコン」

- TSUBAME1同等のエネルギー消費・30倍の電力性能比・PUE=1.28・「Green500」世界一？
  - GPU+マルチコアCPUの大幅活用による高効率化
  - 最先端の冷却法: 密閉型の水冷ラック, 負荷集中での高熱勾配, 夏季の負荷キャップ
  - JST-CREST Ultra Low Power-HPC等の研究成果の応用
- => PUE 1.28以下(他の国内のスパコンセンターは1.7~2.0程度)

3. 「クラウド型スパコン」:総合的学内ITホスティング

- Windows HPC/Linuxなど複数OS, 複数環境のサポート
- 仮想化による種々のデータセンターホスティング機能のサポート
- 教育用/Kioskシステムのバックエンド化、全学アカウント・総合的学内ITの集中化・費用削減

4. 東工大GSICでの種々の基礎研究・メーカー共同開発の成果

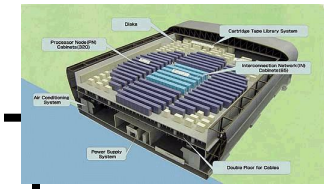
- **JST-CREST "Ultra Low Power HPC",** 科研特定領域「情報爆発」, 文科省-国立情報研 **NAREGI / e-Science**等、多くの基礎研究
- 海外企業とも: **NVIDIA CUDA CoE** (日本初), **Microsoft TCI** 包括共同研究契約
- **NEC, HP, NVIDIA, Microsoft, Voltaire, DDN** 等との共同開発体制

TSUBAME2.0技術パートナーベンダー群

NEC: 主幹・全体設計及びインテグレーション・クラウド管理
HP: ノード開発・全体設計・グリーン; Microsoft: WindowsHPC・クラウド仮想化
NVIDIA: Fermi GPU (ベクトル) CUDA;　Voltaire: QDR Infiniband Network
DDN: 大規模ストレージ; Intel: Westmere & Nehalem-EX CPU (スカラー)
PGI: GPU用ベクトル化コンパイラ
東工大GSIC: 基礎設計・各種GPU技術・スパコン省電力技術・クラスタ運用

TSUBAME2.0の性能向上

地球シミュレータ ⇒ TSUBAME 4年間
30-40倍のダウンサイズ



TSUBAME ⇒
TSUBAME 2.0
4年間で30-40倍

TSUBAME2.0
2.4PF (2010)

Japanese NLP
>10PF(2012)

US >10P
(2011~12?)

U-Tokyo, Kyoto-U,
Tsukuba T2K 61-140TF (2008)

Others

BlueGene/L
360TF(2005)

TSUBAME 1.2
Upgrade
(3Q2008) 170TF

世界一ライン

TSUBAME 1.1 Upgrade 111TF, 1.6
PB, 128GB nodes (1Q2008)

TSUBAME 1.0 80TF (1Q2006)

Earth Simulator
40TF (2002)

Titech
Campus Grid
1.3TF

KEK 59TF
BG/L+SR11100

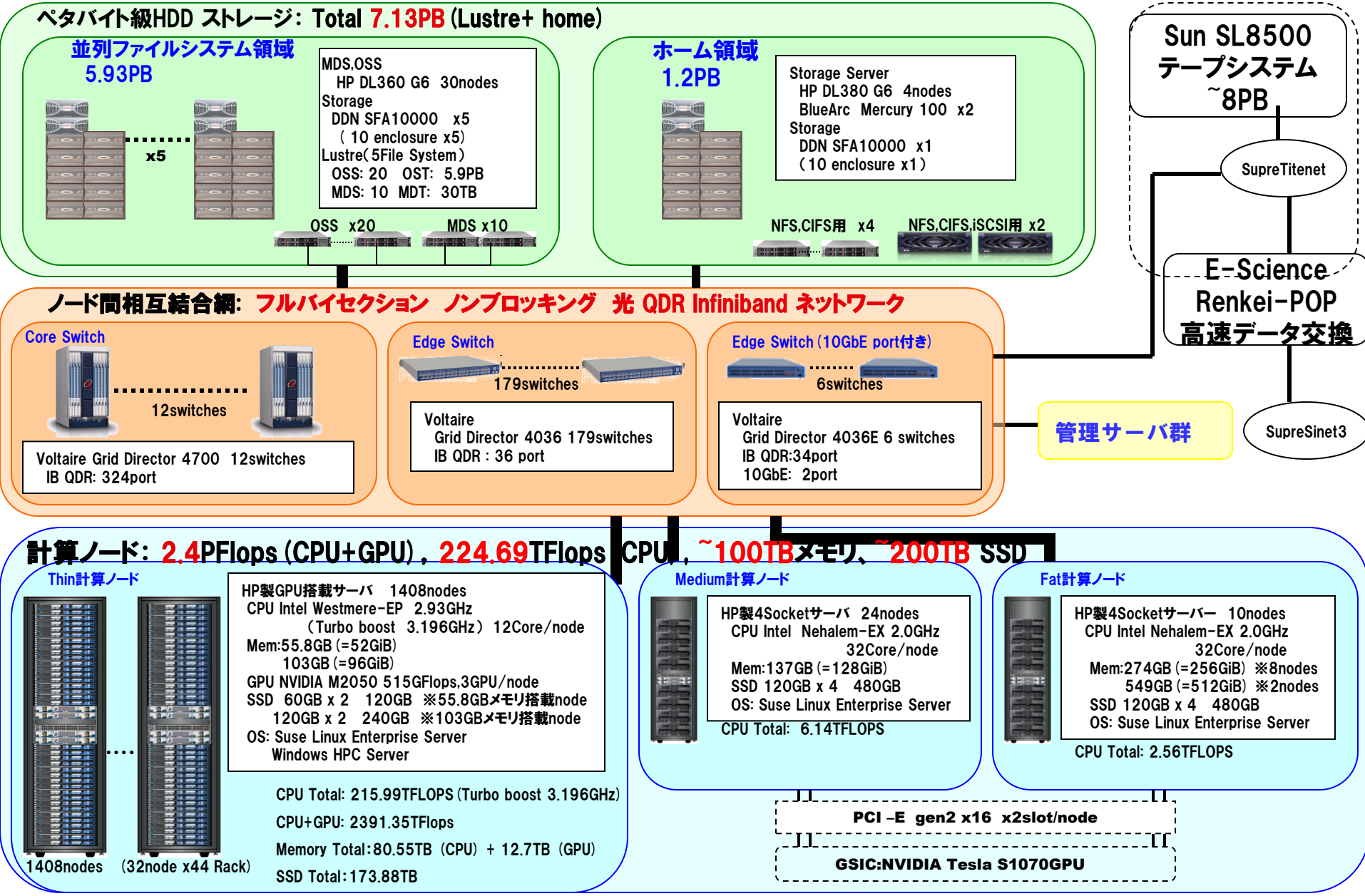
10PF
1PF
100TF
10TF
1TF

2002 2004 2006 2008 2010 2012


Push Here
Do Not
Push


~50 compute racks + 6 switch racks
Two Rooms, Total 160m²
1.4MW (Max, Linpack), 0.48MW (Idle)

TSUBAME2.0 システム概念図



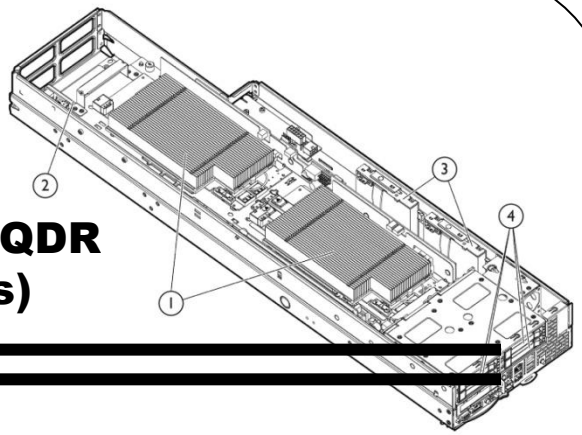
	TSUBAME 1 (2006年, 22億円)	T2K東大 (2008年, 90億円)	TACC Ranger (2008年. 60億円?)	TSUBAME2.0 (2010年, 32億円)
Cores/Node	16	16	16	12(CPU)+1344(GPU)
Node Mem BW(GBytes/s)	20	20	20	64(CPU)+450(GPU)
Node Network BW (Gbps)	20	40	10	80
#Nodes	655	952	3,936	1408(Thin) + 34(Med/Fat)
#Cores (Total)	10,480(CPU)	15,232	62,976	17,664(CPU)+189万(GPU)
# GPUs/Accelerators	360 (ClearSpeed)	0	0	4224 (Tesla M2050)
理論 Peak TFLOPS (倍精度)	80	141	579	2400
合算メモリバンド幅(TB/s) (Flops/Byte)	17 (0.21)	20 (0.13)	80 (0.13)	~720 (0.3)
ネットワークバイセクション(Tbps)	6	41	80	>200
Memory (Tbytes)	21	30	126	100
Linpack (倍精度-TFLOPS)	48	102	433	>1000
合算 3D-FFT 256^3 (TFLOPS)	~13	~20	~80	~700 (GPU only)
HDD Storage (Raw TBytes)	1100	1500	1700	7130
Local SSD Storage/BW (Raw TBytes) (Bandwidth TByte/s)	0/0	0/0	0/0	~200 (0.66PByte/s)
Energy(Incl. Cooling)	850KW/年	~1MW/年	2.4MW Year	~1MW/年
Compute Racks	65	70?	~100	~44

40倍以上

高バンド幅 ベクトル・スカラー混合

TSUBAME2.0 Compute Nodes

Thin Node



Infiniband QDR x2 (80Gbps)

HP SL390G7 (Developed for TSUBAME 2.0)
GPU: NVIDIA Fermi M2050 x 3
515GFlops, 3GByte memory /GPU
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node)
Memory: 54, 96 GB DDR3-1333
SSD：60GBx2, 120GBx2

1408nodes：

4224GPUs：59,136 SIMD Vector Cores, 2175.36TFlops (Double FP)

2816CPUs, 16,896 Scalar Cores: 215.99TFlops

Total： 2391.35TFLOPS

Memory： 80.6TB (CPU) + 12.7TB (GPU)

SSD： 173.9TB

IB QDR

PCI-e Gen2x16 x2
NVIDIA Tesla S1070 GPU

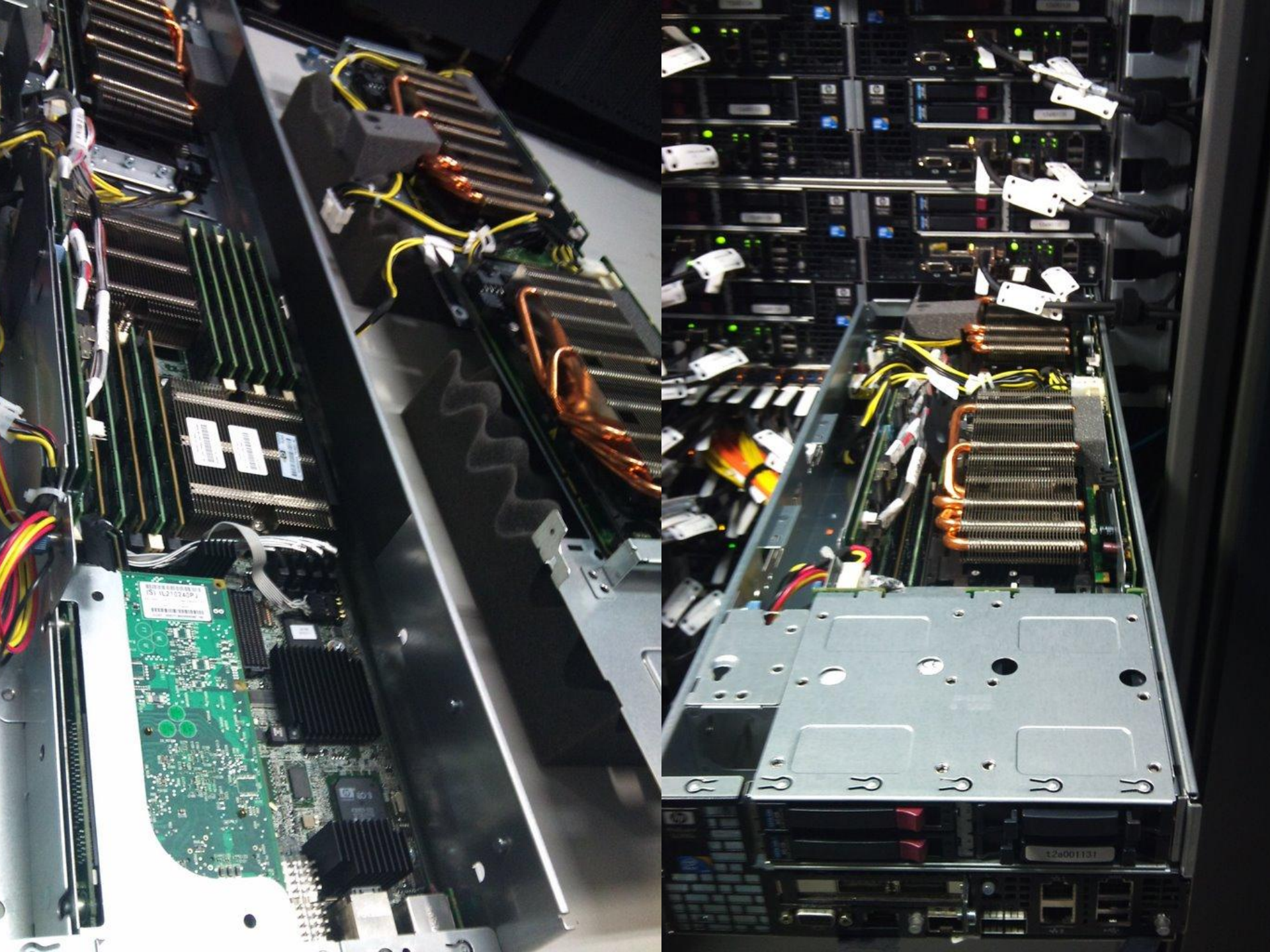
HP 4 Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 (32cores/node)
Memory: 128, 256, 512GB DDR3-1066
SSD：120GB x4 (480GB/node)

34 nodes: 8.7TFlops

Memory: 6.0TB+GPU

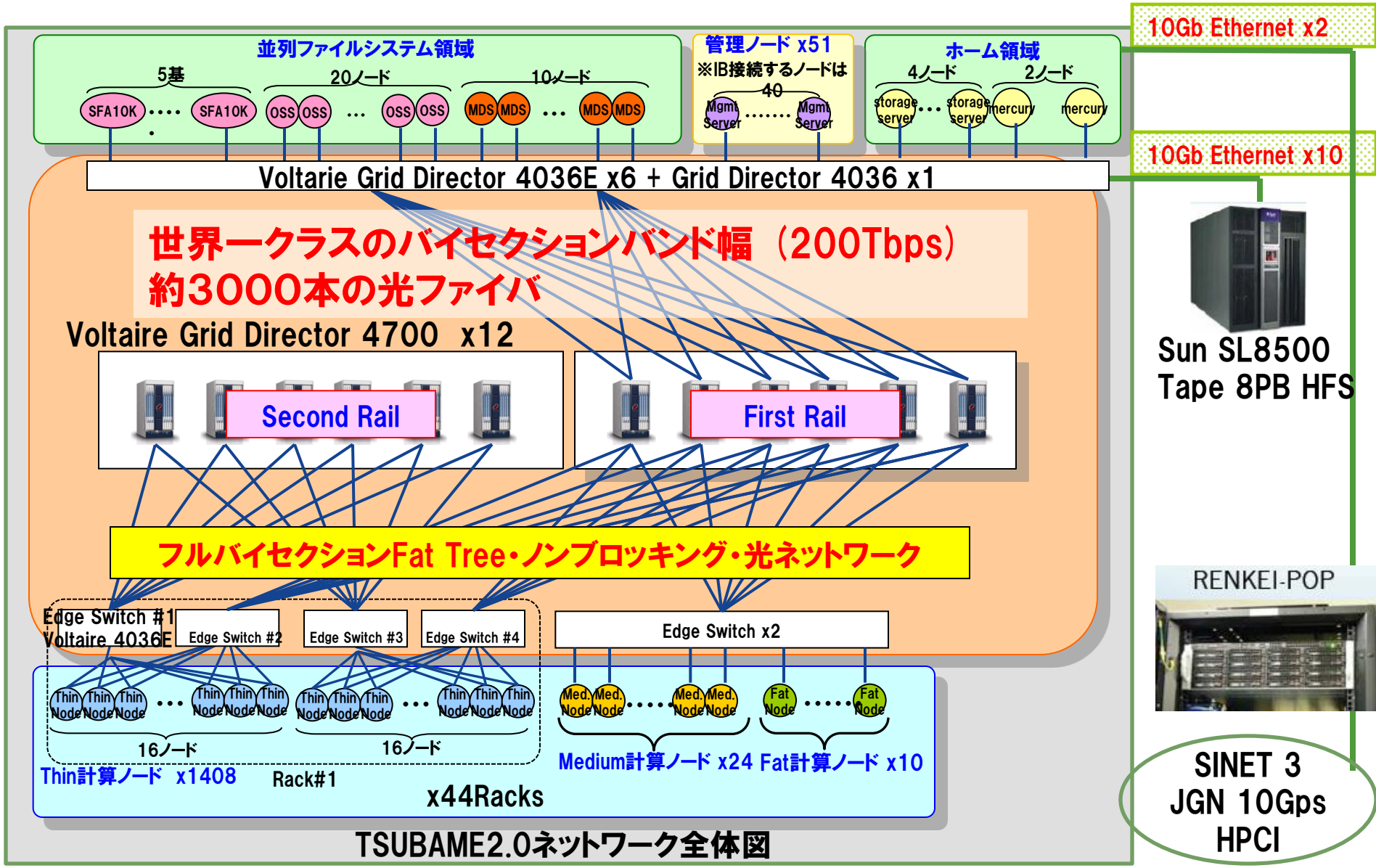
SSD: 16TB+

Total Perf
2.4PFlops
Mem： ~100TB
SSD: ~200TB

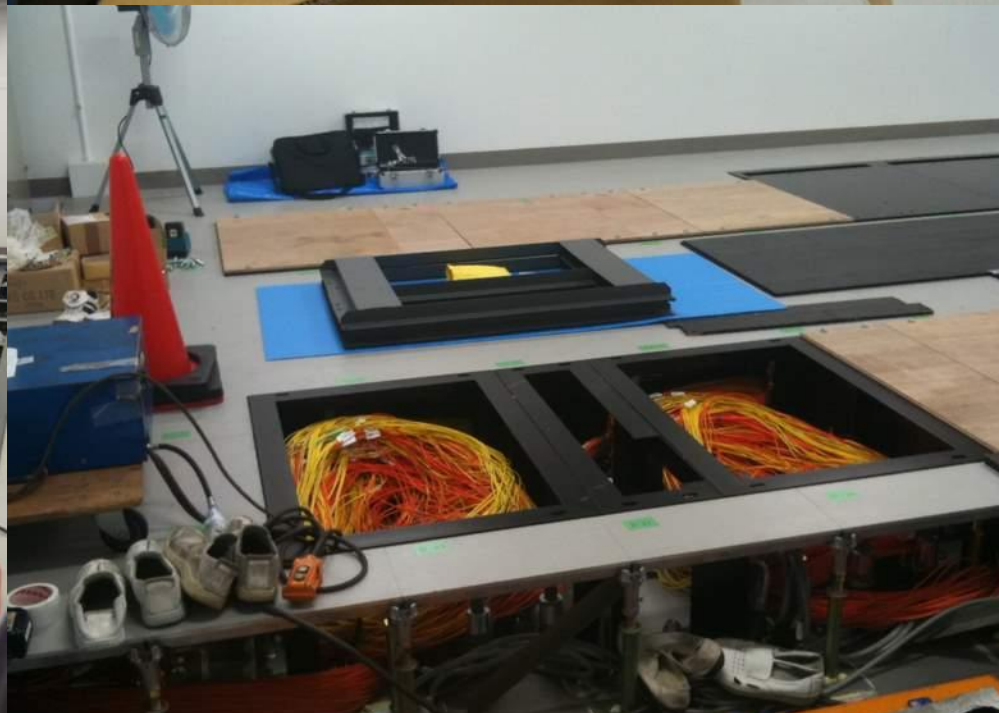



MADE IN TOKYO

TSUBAME2.0ノード間相互結合網



TSUBAME2.0ネットワーク全体図



TSUBAME 2.0ペタバイト級ストレージ

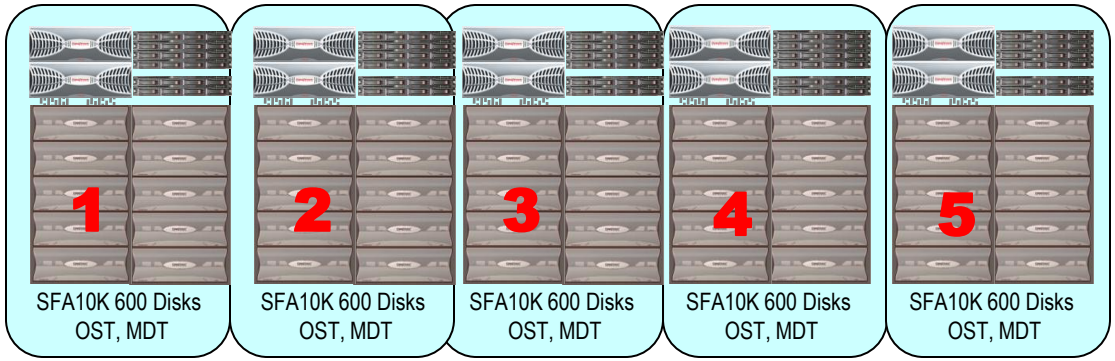
1) 各ノードの短期記憶用SSD、2) Lustre/GPFSを利用した「並列ファイルシステム領域」, NFS,CIFS,iSCSIを備えた「ホーム・クラウドサービス用領域」のHDD群, および 3) 長期保存用テープシステムで構成

Lustre 並列ファイルシステム領域

MDS:HP DL360 G6 x10
- -CPU:Intel Westmere-EP x2 socket (12コア)
- -メモリ:51GB (=48GiB)
- -IB HCA:IB 4X QDR PCI-e G2 x1port

OSS:HP DL360 G6 x20
- -CPU:Intel Westmere-EP x2 socket (12コア)
- -メモリ:25GB (=24GiB)
- -IB HCA:IB 4X QDR PCI-e G2 x2port

ストレージ:DDN SFA10000 x5
- -Total容量:5.93PB
  2TB SATA x 2950 Disks + 600GB SAS x 50 Disks



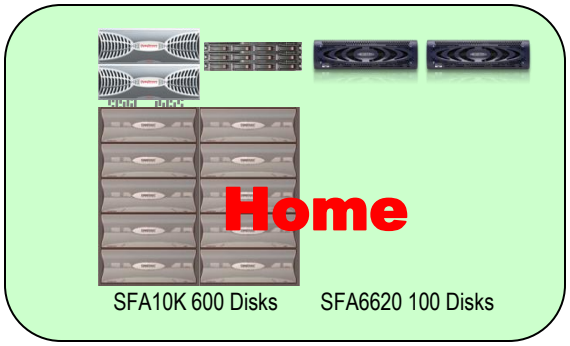
並列ファイルシステム領域 5.93PB

ホーム・クラウドサービス用領域

NFS/CIFS用:HP DL380 G6 x4
- -CPU:Intel Westmere-EP x2 socket (12コア)
- -メモリ:51GB (=48GiB)
- -IB HCA:IB 4X QDR PCI-e G2 x2port

NFS/CIFS/iSCSI アクセラレーション:BlueArc Mercury100 x2
- -10GbE x2

ストレージ:DDN SFA10000 x1
- -Total容量:1.2PB
  2TB SATA x 600 Disks



ホーム領域 1.2PB

約200TB SSD+7.13PB HDD + 約8PBテープ(予定)の大容量階層ストレージ
合算15ペタバイト: 全国大学基盤センター群合算の数倍の容量


11ペタ(10^{15})バイトの
ストレージ

TSUBAME2.0 テープシステム（別調達）
合計15PB以上、階層ファイルシステムの構築

スイッチ（Brocade BI-RX-4-AC）
10G-SR x6ポート
10G-LR x2ポート
1G-Ethernet x24ポート

10GBase-SR(MMF) x6

バックアップ用サーバ x3

SunFire X4275 3nodes
ディスク 300GB(RAID1時)
メモリ 24GB

階層型ストレージ用
サーバ追加予定

FC 4GBps x12 48Gbps

FCスイッチ（Brocade 5300）
FC x80ポート

FC 4GBps x20 80Gbps

100/1000Base-T x1

ACSLSサーバ
（ロボット制御用）

SunFire X4140 1nodes

テープライブラリ SL8500 2台をPTP構成

Sun SL8500 2台
-SL8500 + ストレージエクスパンジョンモジュール3台(5.3PB構成, 6,632スロット)
-SL8500 + ストレージエクスパンジョンモジュール2台(3.9PB構成, 4,904スロット)

本年度~8PB拡張予定

【全体】
容量:4PB(非圧縮時) LTO4 5000巻
(9.2PBまで拡張可能)
ドライブ数:20 drive (128driveまで拡張可能)
ドライブ転送速度:120MB/s(非圧縮時)
スロット数:11,536 slot

GSIC国際棟

20Gbps

SuperTitanet3

既設スイッチ
(Foundry BigIron RX8 Switch)

TSUBAME

Lustre
ファイルシステム領域

ホーム領域
（NFS）

GSIC情報棟

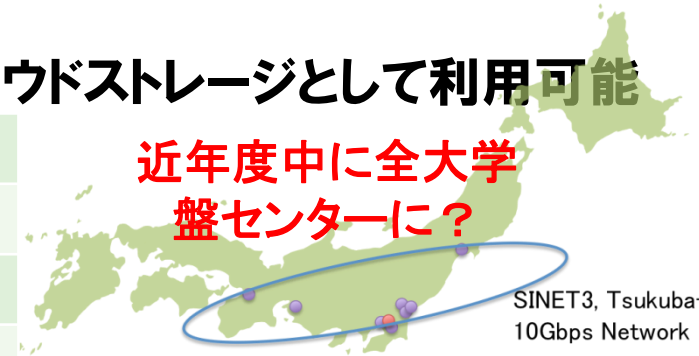
東工大 e-Science RENKEI-POP による分散ストレージ・HPCIへの貢献

- 目的: 高速SINET網を活用・スパコンセンター間データ共有基盤の構築
  - RENKEIプロジェクト(文科省e-Science委託事業)と連携
- ストレージサーバRENKEI-PoP (Point of Presence)の開発・全国に配備
  - 大容量、高速IO性能を備えたデータ転送用サーバアプライアンス
  - SINET3上に広域ファイルシステムGfarm等によりRENKEI-クラウド構築
  - TSUBAME2.0や他の機関のスパコン間の大規模データ交換

CPU	Core i7 975 Extreme (3.33 GHz)
Memory	12GB (DDR3 PC3-10600 , 2GB*6)
NIC	10GbE (without TCP/IP Offload Engine)
System Disk	500GB HDD
SSD RAID	30TB (RAID 5, 2TB HDD x 16)

- 現在9拠点に配備、110TBの高速分散クラウドストレージとして利用可能

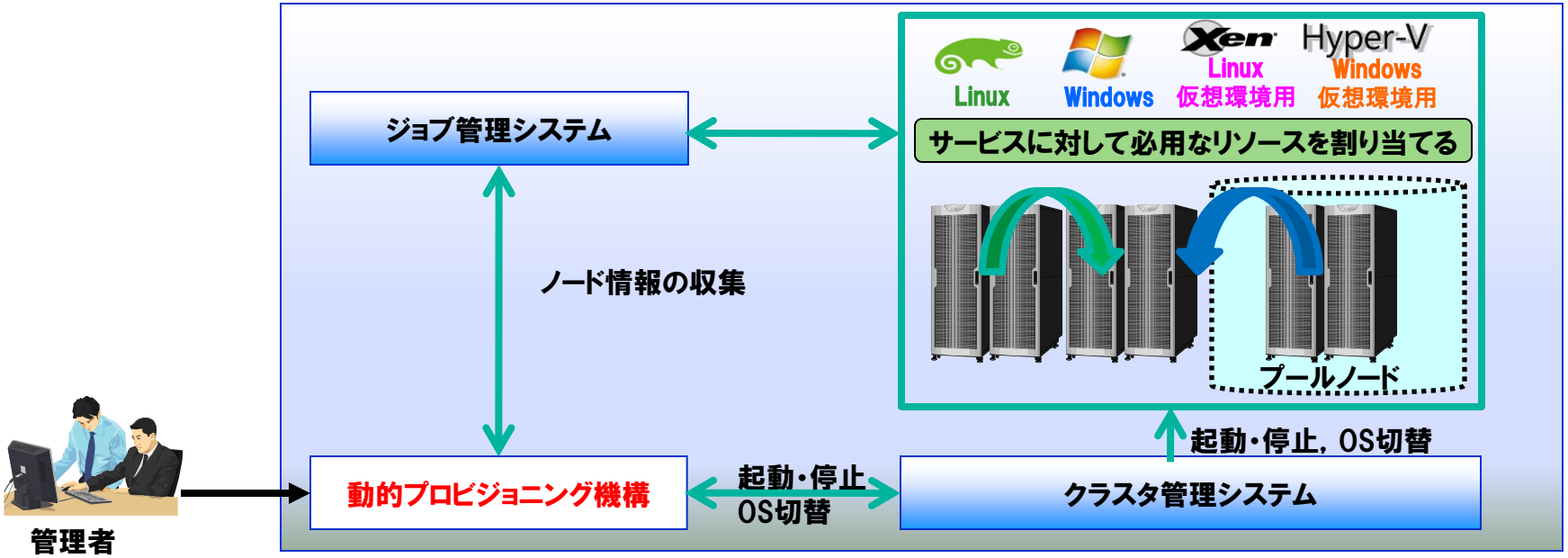
東京工業大学	大阪大学
国立情報学研究所	高エネルギー加速器研究機構
名古屋大学	筑波大学
産業技術総合研究所	東北大学



TSUBAME2.0クラウド運用形態

OSを動的に変更する「動的プロビジョニング機構」

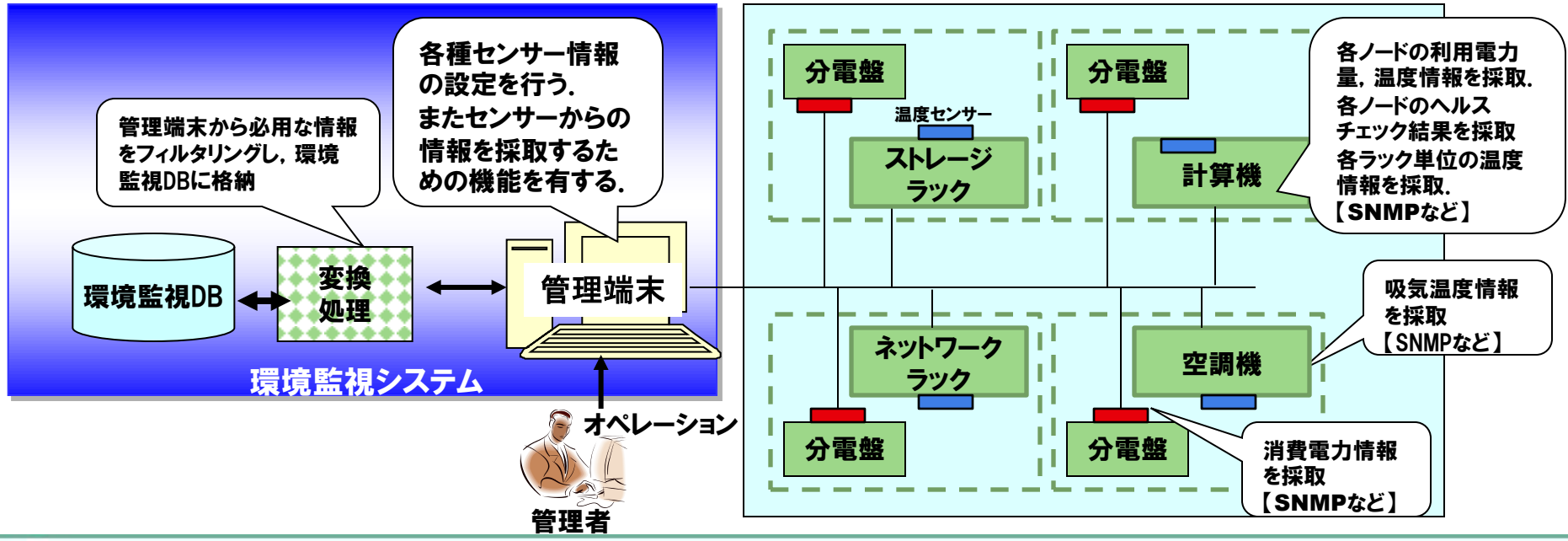
- 「ジョブ管理システム」「クラスタ管理システム」と連携
- プールノード(サービスに割り当てられていない計算ノード)を利用し，余剰の計算リソースをサービスに割り当て
- Linux,Windows双方のバッチスケジューラにより計算ノードを協調管理
- Linuxノード，Windowsノードの動的な増減が可能
- バーチャルマシン上の仮想計算ノードもスケジュール可能な資源として動的にバッチスケジューラの管理対象



グリーンスパコン: 環境監視システム

各計算ノード，ラック，及び計算機室の温度情報・消費電力等を監視する「環境監視システム」.

- センサー情報及び各計算ノードの情報をオンラインでモニタリング
  - 温度情報（温度センサーから取得）
  - ヘルスチェック結果，サービス提供状況，故障の有無
  - 消費電力（各ノード・及び各分電盤から取得）

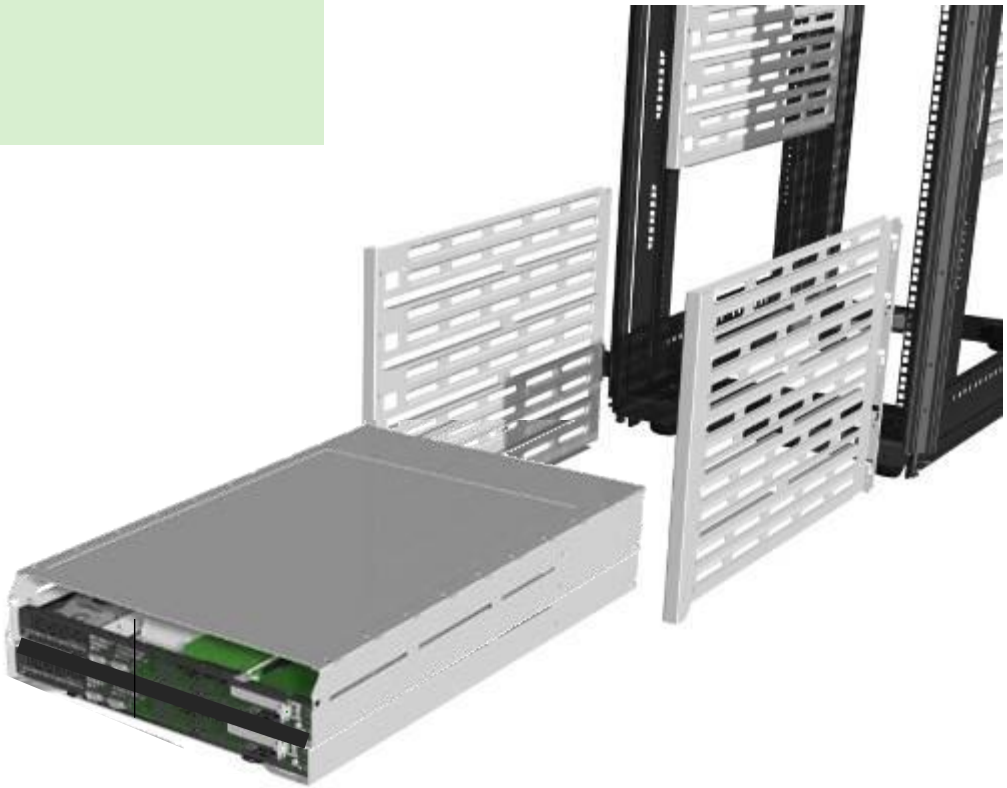


GPU-optimized node in HP Skinless Packaging

Typical Tsubame2 node:

- 2 Intel® Xeon® Processor X5670 (six-core 2.93Ghz)
- 54 GB or 96 GB memory
- 2 SSDs per node
- Dual Rail IB
- 3 NVIDIA M2050 GPU board

- Nodes go into chassis with shared fans and power
- Chassis mounted into lightweight racks
- Easy assembly
- Lower weight
- Reduced heat retention
- Modular and flexible
- Standard 19” racks



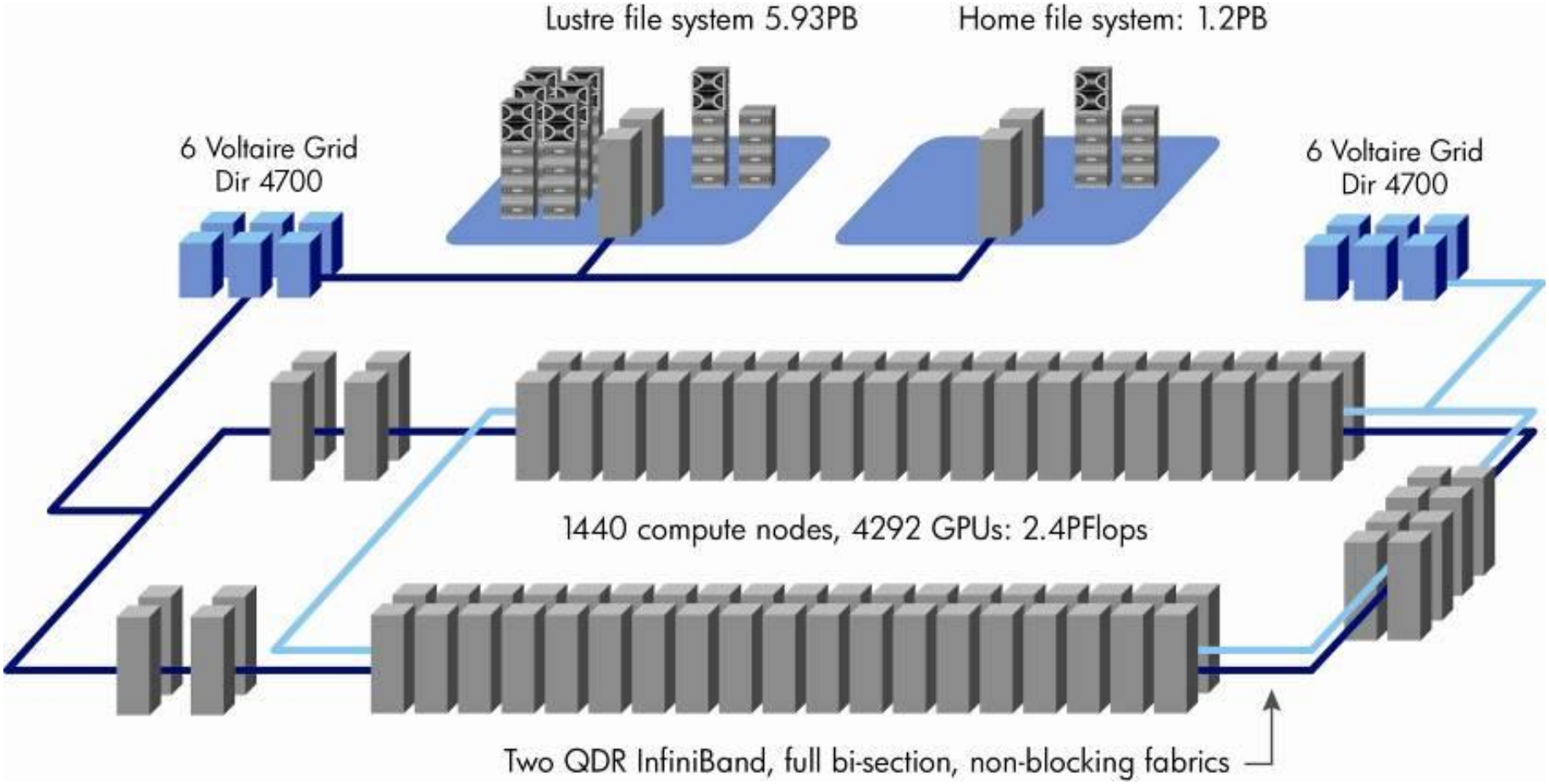
Compute Rack Building Block for Tsubame2

HP Modular Cooling System G2 rack

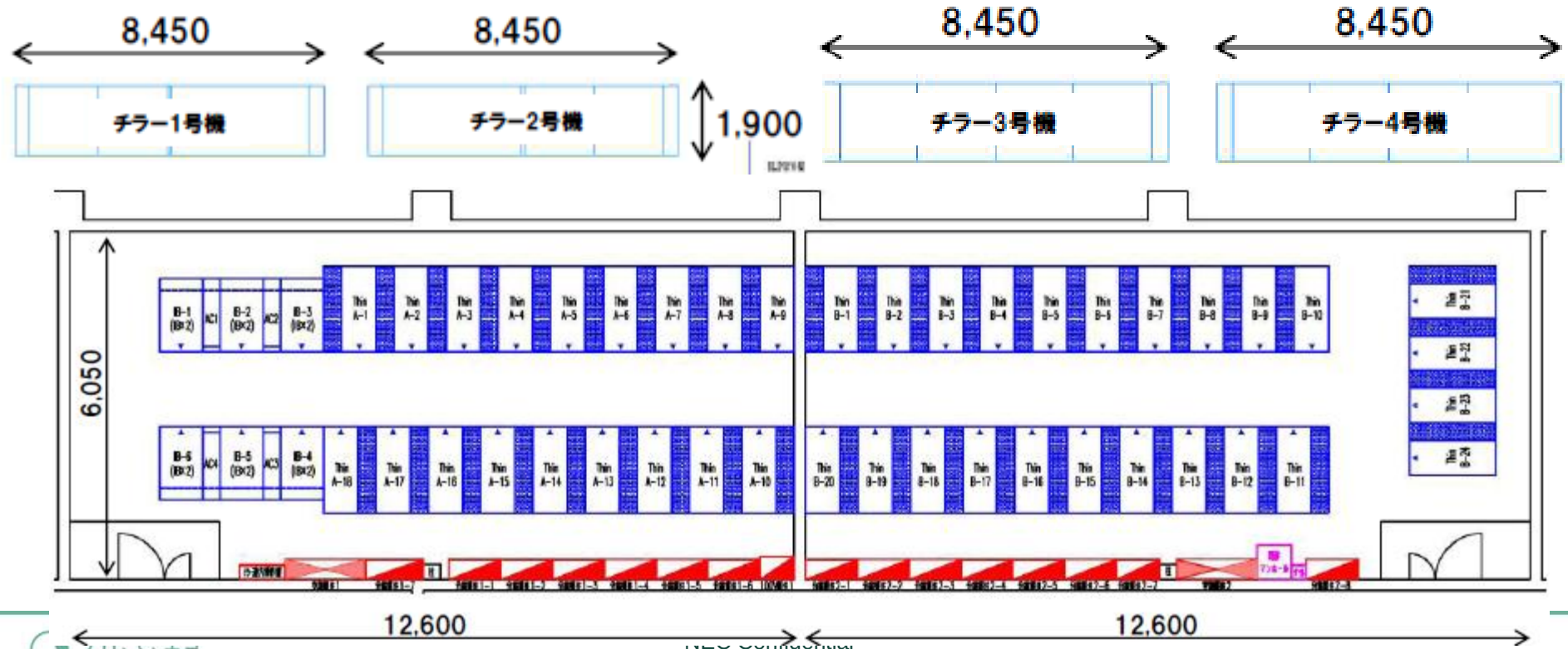
- 42U HP Modular Cooling System G2 rack
- 30 nodes per rack – 90 GPUs
- 8 chassis with Advanced Power Management
- 1 HP Network Switch for shared console and admin local area network
- 2 Airflow Dam (Liquid Cooled)
- 4 Voltaire 4036 Leaf Switch
- Power distribution units

Power per rack approximately **35KW**

Pulling It All Together

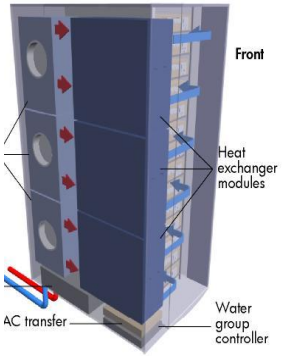


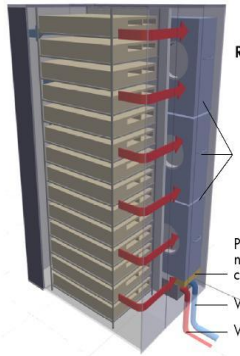
TSUBAME2.0 レイアウト
(全体で200m2程度, TSUBAME1の2/3以下)



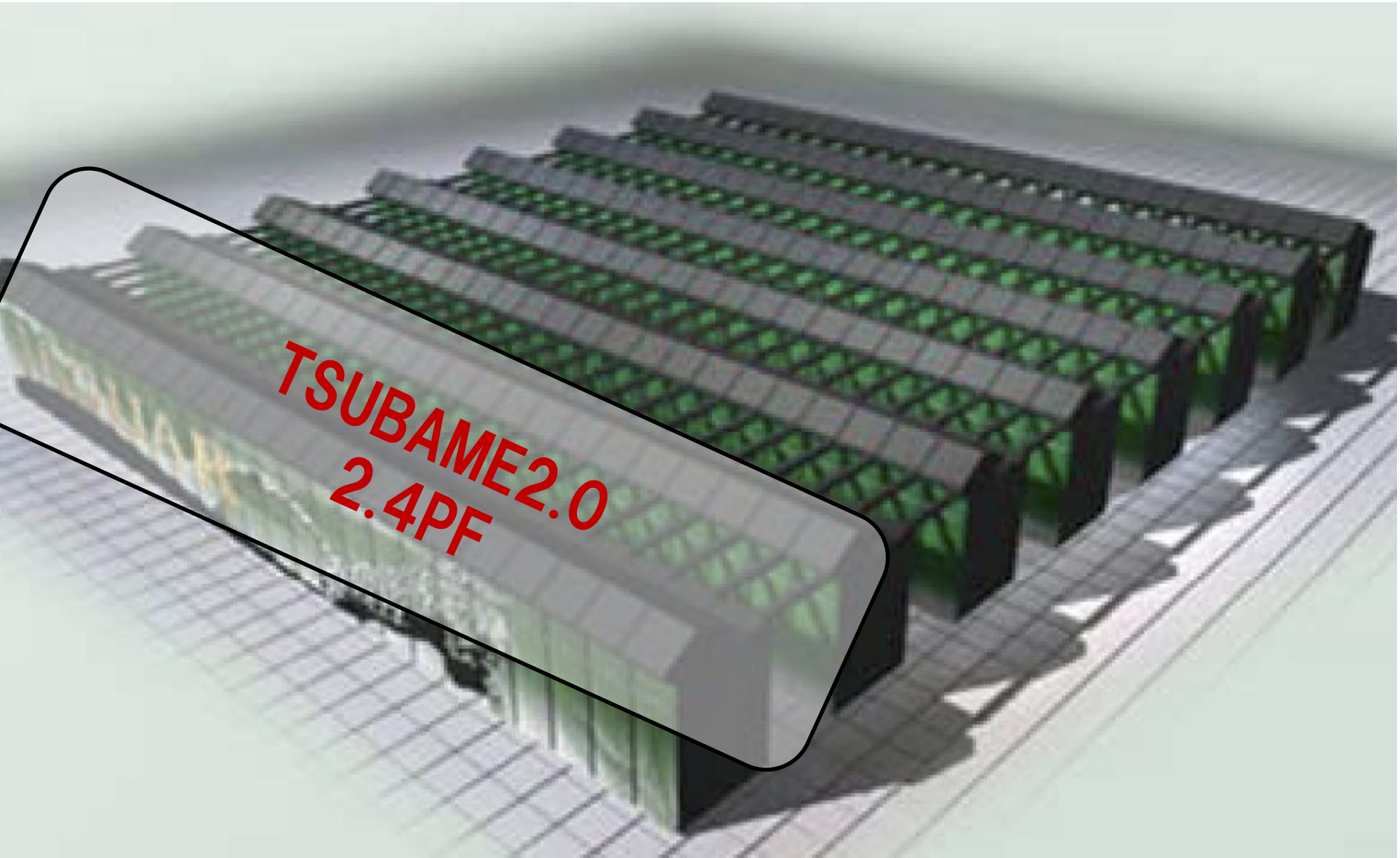
グリーンスパコン：HP Modular Cooling System G2 による高密度実装・水冷キャビネット冷却

- ラック内に熱交換システムを内蔵した密閉型水冷システム
- 高密度な冷却が可能・ラックあたり最大35kW（世界最高）
  通常のデータセンターの10倍!!
- サーバの吸入口に均質な冷却風を提供
- ドア平開は自動化・加湿不要
- 完全自動温度制御による最適な消費電力点の制御
- 95% から 97% の熱を水冷で除去
- ポリカーボネート製のドアにより大幅なノイズ削減



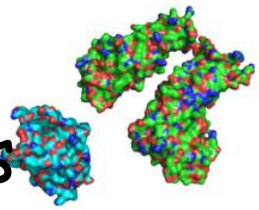


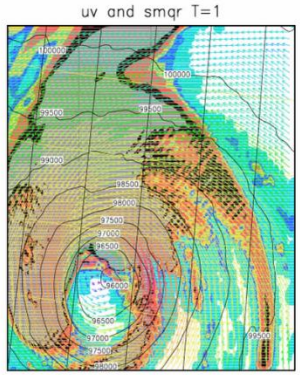
ORNL Jaguar and Tsubame 2.0
Similar Peak Performance, 1/4 the Size and Power



TSUBAME2.0 アプリケーション性能予測

- 1.192 TFlops Linpack [IEEE IPDPS 2010]
  - Top ranks Green 500?
- ~0.5 PF 3D Protein Rigid Docking (Node 3-D FFT) [SC08, SC09]
- 145Tlops ASUCA Weather Forecast [SC10 Best Student Paper Finalist]
- Multiscale Simulation of Cardiovascular flows [SC10 Gordon Bell Finalist]
- Various FEM, Genomics, MD/MO (w/IM Apps: search, optimization, …





TSUBAME2.0世界ランキング
スパコン二大リスト (2010年11月)

The Top 500 (ベンチマーク絶対性能、ペタフロップス)

1位: 2.566 中国防衛大 Tianhe 1-A (11)
2位: 1.758 :米国オークリッジ国立研究所 Cray Jaguar (81)
3位: 1.271 :中国深圳国立スパコンセンター Dawning Nebulae (13)
4位: 1.192 :日本 東工大/HP/NEC TSUBAME2.0 (2)
5位: 1.054 :米国ローレンスバークレー国立研究所 Cray Hopper (30)
6位: 1.050 :仏CEA国立研究所 Bull Bullx (97)
7位: 1.042 :米国オークリッジ国立研究所 IBM Roadrunner (16)
33位(日本2位): 0.1914:日本原子力研究開発機構/富士通 (95)

(Green500 rank)

The Green 500 (ベンチマーク電力性能、メガフロップス/W)

1位: 1684.20 : 米国 IBM研究所 BlueGene/Q プロトタイプ (116)
2位: 958.35 :日本 東工大/HP/NEC TSUBAME2.0 (4)
3位: 933.06 :米国 NCSA Hybrid Cluster実験機 (403)
4位: 828.67 : 日本 理研 京 (170)
5-7位: 773.38 : ドイツ ユーリッヒ大等 IBM QPACE SFB TR (207-209)
10位(日本3位): 636.36 :日本 環境研 (102)

(Top500 rank)

"Little Green 500" では TSUBAME2.0の実験構成が
1.037 Gigaflops/W 達成 (米Microsoftとの共同研究)


THE GREEN 500™
sponsored by
SUPERMICRO
This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.
GSIC Center, Tokyo Institute of Technology
Is recognized as the
Greenest Production Supercomputer in the World
on the world's Green500 List of computer systems as of
November 2010
Wu-chun Feng, Co-Chair
Kirk Cameron, Co-Chair

GPUで最速スパコン

高速計算 開発費も安く

中国のスーパーコンピューターが計算速度で初めて世界一になったスパコンランキング「TOP500」では、中央演算処理装置（CPU）に加え、「GPU」＝■■＝と呼ばれる画像処理装置を多く積んだスパコンが、上位5台のうち3台を占めた。高速の割に小型で消費電力も少ないのが特徴だ。スパコンは大規模化が限界に来ており、GPUや専用演算素子を駆使した新世代へと、急速に多様化が進んでいる。（東山正宣）

世界一の「天河1号A」（中国国防科学技術大）は、2位の「ジャガー」（米オークリッジ研）の約1・5倍の計算速度をたたき出した。

天河の部品のほとんどは米国製だ。TOP500などによると、米インテル製のCPU1万4千個と、米エヌビディア製のGPU7千個からなる。これまでのほとんどのスパコンがCPUだけを使っていたのに対して、天河はCPUをCPU代わりに使っているのが特徴だ。

GPU

Graphics Processing Unit（画像処理装置）の略。コンピューターの画面に写真や動画を表示する装置。ゲームの3D表示やハイビジョン当たり前になった2000年代半ばから急速に高性能化が進んだ。最新の機種では数千本の中に512個

本職は画像処理

3位の星雲（中国深圳スパコンセンター）や4位のツバメ2・0（東工大）も構成や部品はほとんど同じ。3台の計算速度の違いはCPUとGPUの数の違いを反映している。東工大の松岡聡教授は「数を増やしても、使いこなして性能を出すのは簡単ではない。中国もなかなかやるようになってきた」と話す。

GPUの本来の仕事である画像処理では、簡単な計算を数多くこなす必要がある。このため、GPUには単純な計算を担う小さな演算素子がたくさん詰め込まれている。一方、CPUは条件ごとに場合分けをするなど複雑な計算をさせるため、それに対応した大型の素子を積んでいる。二つを使い分け、計算をうまく割り振れれば効果は絶大だ。

エヌビディアによると、天河の計算速度（1秒間に2566兆回）の8割をGPUが担っているという。同社日本法人の林憲一マーケティングマネジャーは「CPUだけで同じ計算速度を出そうとすると5万個必要になる。床面積は倍、消費電力は3倍になっていただろう」と試

開発費も安く

市販品を使っているので開発費も安い。天河は約80億円、ツバメは30億円だった。2002年に世界一をとった「地球シミュレータ」が600億円、理化学研究所が建設している「京」が、建屋も含めて1100億円かかるのに比べ、けた違いに安い。

市販化も早い。ツバメを開発したNECはすでに部品ごとのリース販売を始めた。ツバメは大型冷蔵庫ほどのラック50台からなる。その1台ずつが地球シミュレータと同等の計算速度を持つ。02年から8年間で大きさも価格も千分の1になった。

クラウドで普及、省エネに貢献

計算速度を競う「TOP500」と並んで注目されるコンテストが、18日に発表される「GREEN500」だ。TOP500に選ばれたスパコン500台が消費電力あたりの計算速度を競う。2007年に始まった。

松岡教授によると、世界の電力の1割程度がIT（情報技術）関連で使われ、そのうちスパコンの消費電力は4～5％程度。この割合は上がると予想される。クラウドコンピューティングを通じてスパコンが普及するからだ。

クラウドコンピューティングでは、計算を手元のパソコンではなく、ネットの動画閲覧サイト「ユーチューブ」が代表的だ。膨大な数のパソコンから性能された計算を、スパコンが担う流れが進んでいる。

省エネ性能の進歩は、スパコンがパソコンを上回る。ツバメ2・0の先代機種は、同じ計算をするのに必要な電力が、パソコンよりも多かったが、今は効率が20倍上がり、パソコンの3分の1になった。クラウド化が進めばIT全体の電力消費が減らせると、松岡教授はみる。

GPUを使ったツバメ2・0は、GREEN500で上位入りが確実視されている。

（http://aspara.asahi.com）の「aサロン・科学面にようこそ」にもトップ記事を掲載しています。

朝日20101119

省エネスパコン 日本2位

東工大「ツバメ」、理研4位

スーパーコンピューターの省エネ性能を競う世界ランキングが18日に発表され、東京工業大の「ツバメ2・0」が2位、理化学研究所が神戸に建設している「京」が4位になった。1位は米IBMが開発中の「ブルージーンQ」、3位は米国立スーパーコンピューター応用研究所の試験機。すでに運用されているスパコンとしてはツバメが世界一だった。

ランキングは「グリーン500」で消費電力当たりの計算速度を競う。1ワット当たりの計算速度は、ブルージーンが毎秒16億8400万回、ツバメが9億5800万回。ツバメは、計算速度を競う世界ランキング「TOP500」では4位だった。京は現在、全体の0・5%しかできていないが、高性能が示された。

スパコンは消費電力の問題で大型化が限界に近づいている。ブルージーンやツバメは従来の演算装置だけに頼らない新世代のスパコンだ。（東山正宣、ニューオーリンズ＝小宮山亮磨）

朝日20101119(夕刊)

スパコン省エネ性能

東工大、世界2位

日本製 10位内に3つ

米バージニア工科大学が集計したスーパーコンピューターの省エネ性能ランキング「グリーン500」で、日本製のスパコンが10位内に3つ入った。実際に運用しているスパコンの中では東工大のTSUBAME2・0が1位。消費電力を抑えやすい画像処理チップ（GPU）を採用したほか、冷却機構を工夫するなどして省エネ性能を高めた。

東京工業大学のスパコン「TSUBAME2・0」が世界2位となった。日本勢では、次世代スパコン「京（けい）」の試作機が4位、国立環境研究所のスパコンは10位に入った。

グリーン500は消費電力当たりの演算能力の高さを競う。1位は米IBM製のスパコンだが、同機種は試作機の扱い

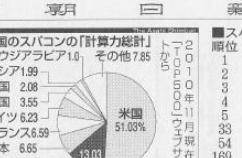
日経20101121

中国スパコン 軍の影

計算速度 世界一

米、技術力を注視

日本、



■スパコン世界ランキング

順位	国	名称	設置機関	計算速度
1	中	天河1号	国防科学技術大	2566
2	米	ジャガー	オークリッジ研究所	1759
3	中	星雲	深圳スパコンセンター	1271
4	日	ツバメ2.0	東京工業大	1192
5	米	ホッパー	ローレンス・バークリー研究所	1054
33	日	BX900	日本原子力研究開発機構	191
54	日	地球シミュレータ	海洋研究開発機構	122
169	日	京（開発中）	理化学研究所	48

米「TOP500」コンテスト（2010年11月）から。計算速度は兆回／秒

朝日20101118

ペタフロップス？ギガフロップス/W?

6.6万倍高速

3倍省エネ

<<

4.4万倍データ

Laptop: SONY Vaio type Z (VPCZ1)
CPU: Intel Core i7 620M (2.66GHz)
MEMORY: DDR3-1066 4GBx2
OS: Microsoft Windows 7 Ultimate 64bit
HPL: Intel(R) Optimized LINPACK Benchmark for Windows (10.2.6.015)
256GB HDD

18.1 ギガ(10^9)フロップス

369 メガ(10^6)フロップス / Watt

Supercomputer: TSUBAME 2.0
CPU: 2714 Intel Westmere 2.93 Ghz
GPU: 4071 nVidia Fermi M2050
MEMORY: DDR3-1333 80TB + GDDR5 12TB
OS: SuSE Linux 11 + Windows HPC Server R2
HPL: Tokyo Tech Heterogeneous HPL
11PB Hierarchical Storage

1.192 ペタ(10^{15})フロップス

1037 メガ(10^6)フロップス / Watt

TSUBAME2.0へのJST-CREST Ultra Low Power HPCの成果の反映

- GPU中心の2.4 PF我が国最速・世界3~4位のスパコン
  - 1432 nodes, Intel Westmere/Nehalem EX
  - 4224 NVIDIA Tesla (Fermi) M2050 GPUs
  - ~76,600 total CPU and GPU "cores", High Bandwidth
  - 電力効率世界一？(後述)
- 省電力運用のための仕組みが随所
  - ノード・ラック・配電版など随所の電力センサーネットワーク
  - 大量の温度センサー(ノード18個=>全体で2万個以、ファンセンサー…)
  - ノード単位の電力キャップ・高効率冷却ノード設計
  - 密閉型水冷ラック/チラー (35KW => PUE=1.28以下)
- 今後ULP-CRESTの成果の適用
  - 開発された高性能・省電力GPUライブラリ等の利用
  - 省電力自動チューニング
  - 省電力スケジューリング(温度感知=>マイグレーション)

ULPHPCにおける
省電力化の研究
(サンプル)

Measuring GPU Power Consumption

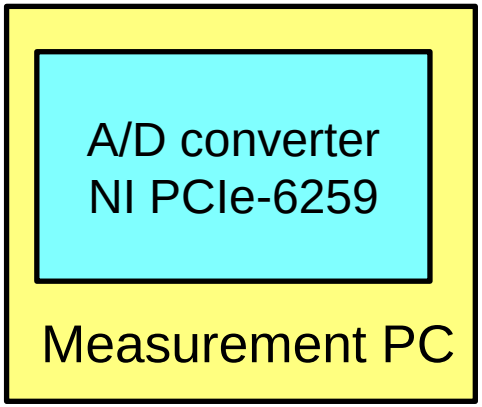
- Two power sources
  - Via PCI Express: < 75 W
  - Direct inputs from PSU: < 240 W
- Uses current clamp sensors around the power inputs

Precise and Accurate Measurement with Current Probes

Attaches current sensors to two power lines in PCIe

+

Direct power inputs from PSU



Reads currents at 100 us interval

統計的GPU電力モデリング [IEEE IGCC10]

- GPUの消費電力を統計的に推定
- 性能プロファイル（パフォーマンスカウンタ）を説明変数とした線形回帰モデル

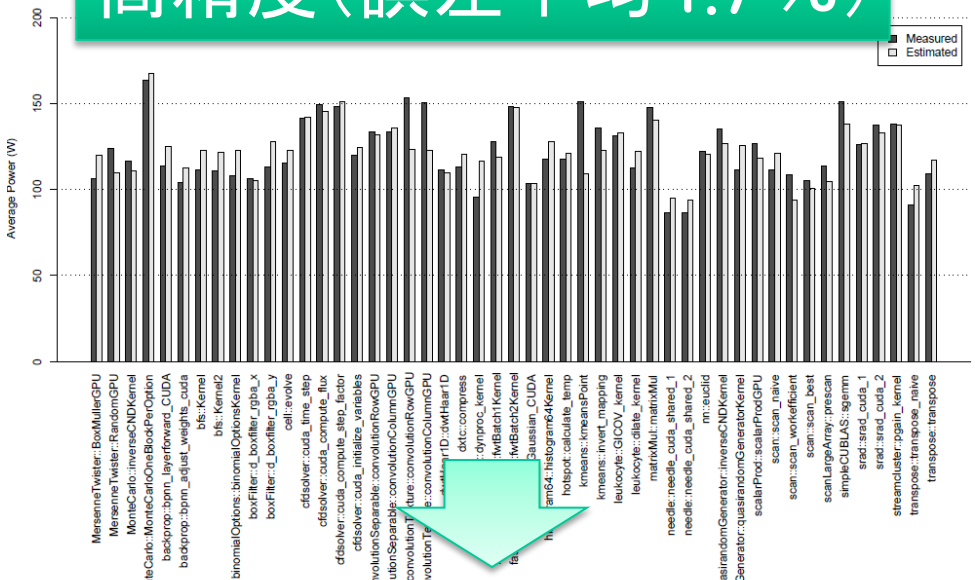
GPUパフォーマンスカウンタ

$$p = \sum_{i=1}^{n} \alpha_i c_i + \beta$$

平均消費電力

- リッジリグレッションによる過学習の防止
- クロスフィッティングによる最適パラメータの決定

高精度（誤差平均4.7%）



今後：電力モデルによる電力最適化

線型モデルでも十分な精度

↓

億単位のプロセッサからなる**エクサスケール**における最適化の実現可能性

Actual vs. Estimated Power

4.7% error in

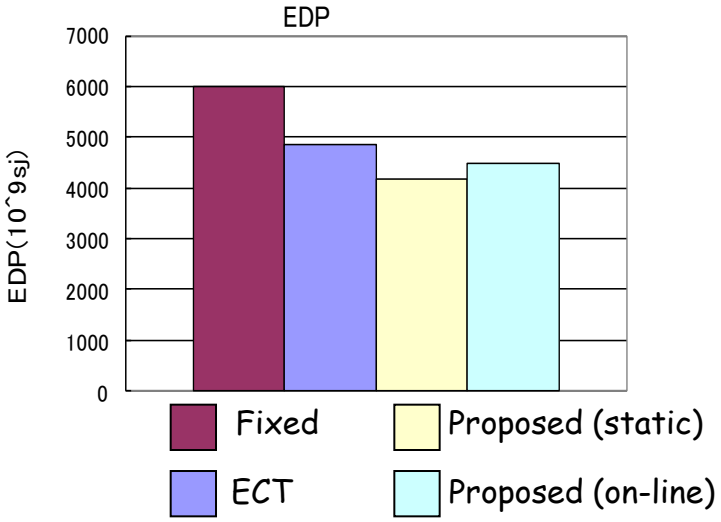
Low Power Scheduling in GPU Clusters

[IEEE IPDPS-HPPAC 09]

Objective: Optimize CPU/GPU Heterogeneous

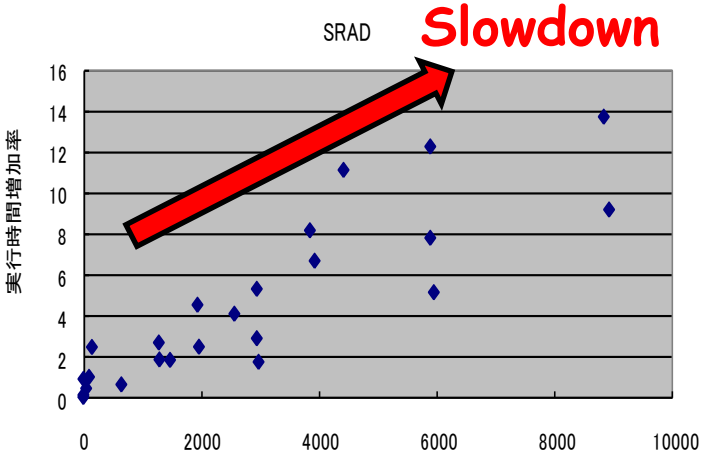
- Optimally schedule mixed sets of jobs executable on either CPU or GPU but w/different performance & power
- Assume GPU accel. factor (%) known

30% Improvement Energy-Delay Product

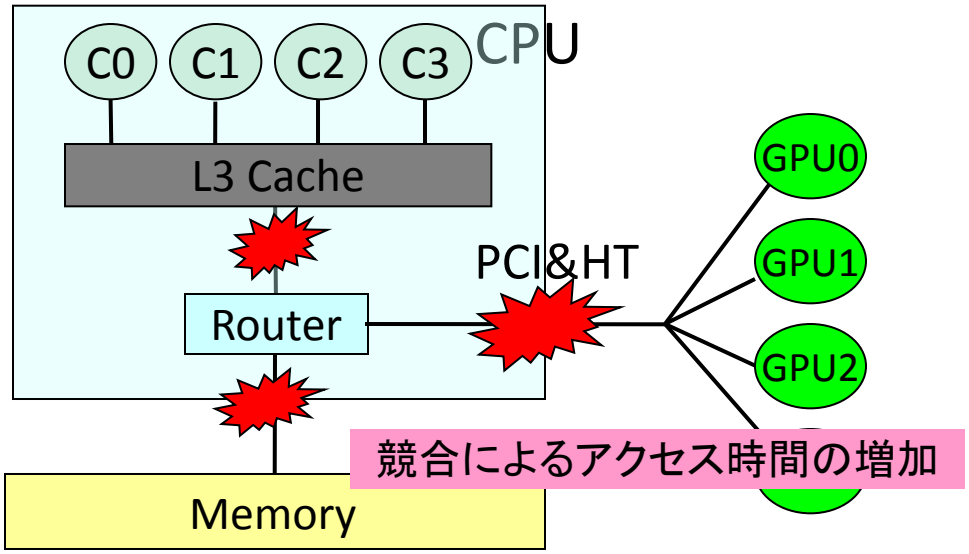


TODO: More realistic environment

- Different app. power profile
- PCI bus vs. memory conflict
  - GPU applications slow down by 10 % or more when co-scheduled with memory-intensive CPU app.



GPUクラスタにおける省電力タスクスケジューリング
(ノード内のプロセス間のメモリやPCIバス競合を考慮)



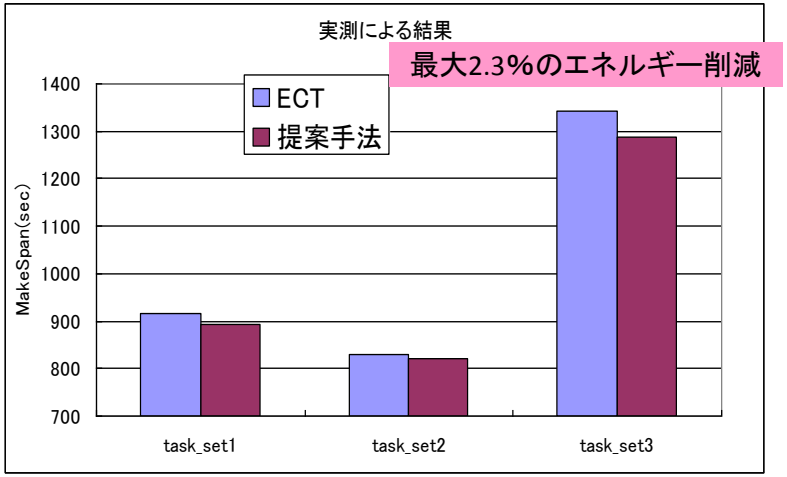
各プロセスは以下のいずれかの状態にある.
(1)メモリアクセス
(2)PCI転送中(双方向)
(3)コア内で計算中
プロセスの時間増加率＝各状態にある確率r

$$\sum_{s0}\sum_{s1}\sum_{s2}\sum_{s3}\alpha(s0,\{s1,s2,s3\})r_0(s0)r_1(s1)r_2(s2)r_3(s3)$$

ECT(Earliest Complete Time)スケジューリングでプロセス間の競合を加味した手法を提案。

各ジョブの下記情報(単独実行時)を既知と仮定.
- 実行時間
- メモリアクセス量(パフォーマンスカウンタから取得)
- PCI転送量(CUDAプロファイラから取得)

これらからプロセス実行時間の増加率をモデルにより推定



TSUBAMEにおける次世代気象予測（気象庁との共同研究）

青木G（東工大）

メソスケール大気モデル:

雲の解像： 3次元非静力学平衡モデル

Compressible equation taking consideration of sound waves.

メソスケール

2000 km

台風

a few km

竜巻, ダウンバースト
集中豪雨

Next Generation Numerical Weather Prediction[SC10]

Collaboration: Japan Meteorological Agency

Meso-scale Atmosphere Model:

Cloud Resolving Non-hydrostatic model

ex. WRF(Weather Research and Forecast)

Typhoon ~ 1000km

1~ 10km
Tornado,
Down burst,
Heavy Rain

WSM5 (WRF Single Moment 5-tracer) Microphysics*

Represents condensation, precipitation and thermodynamic effects of latent heat release

1 % of lines of code, 25 % of elapsed time

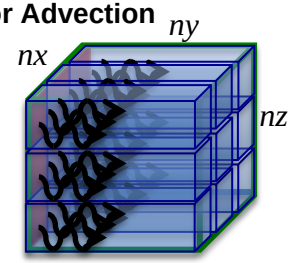
⇒ 20 x boost in microphysics (1.2 - 1.3 x overall improvement)

ASUCA : full GPU Implementation

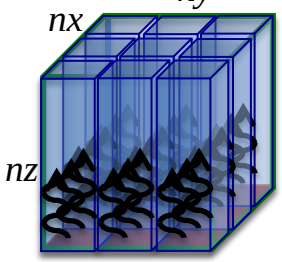
developed by Japan Meteorological Agency

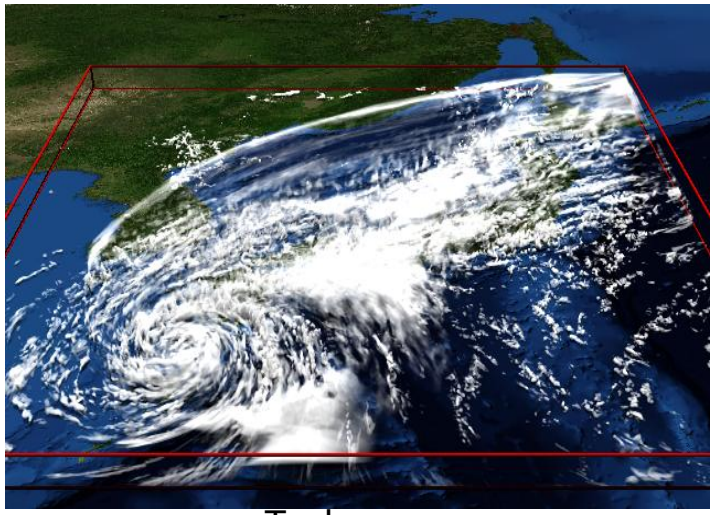
TSUBAME 2.0 : 145 Tflops
World Record !!!

Block Division for Advection

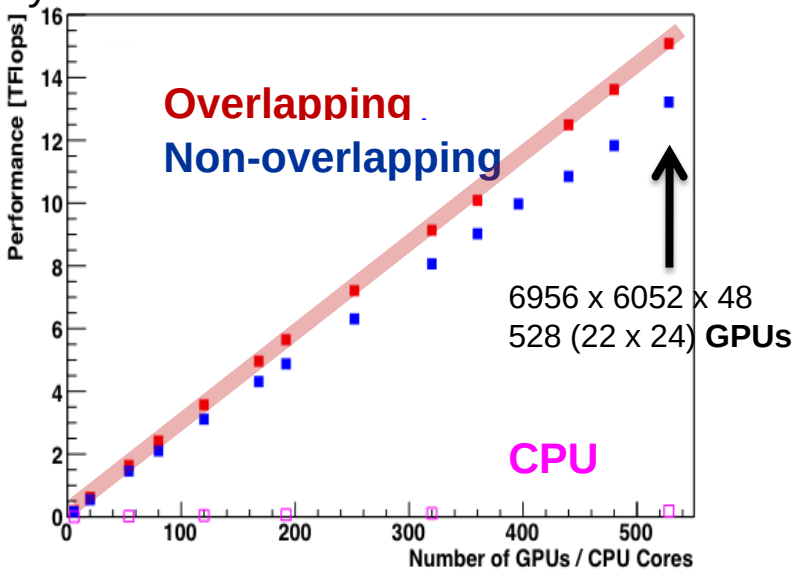


for 1-D Helmholtz eq.

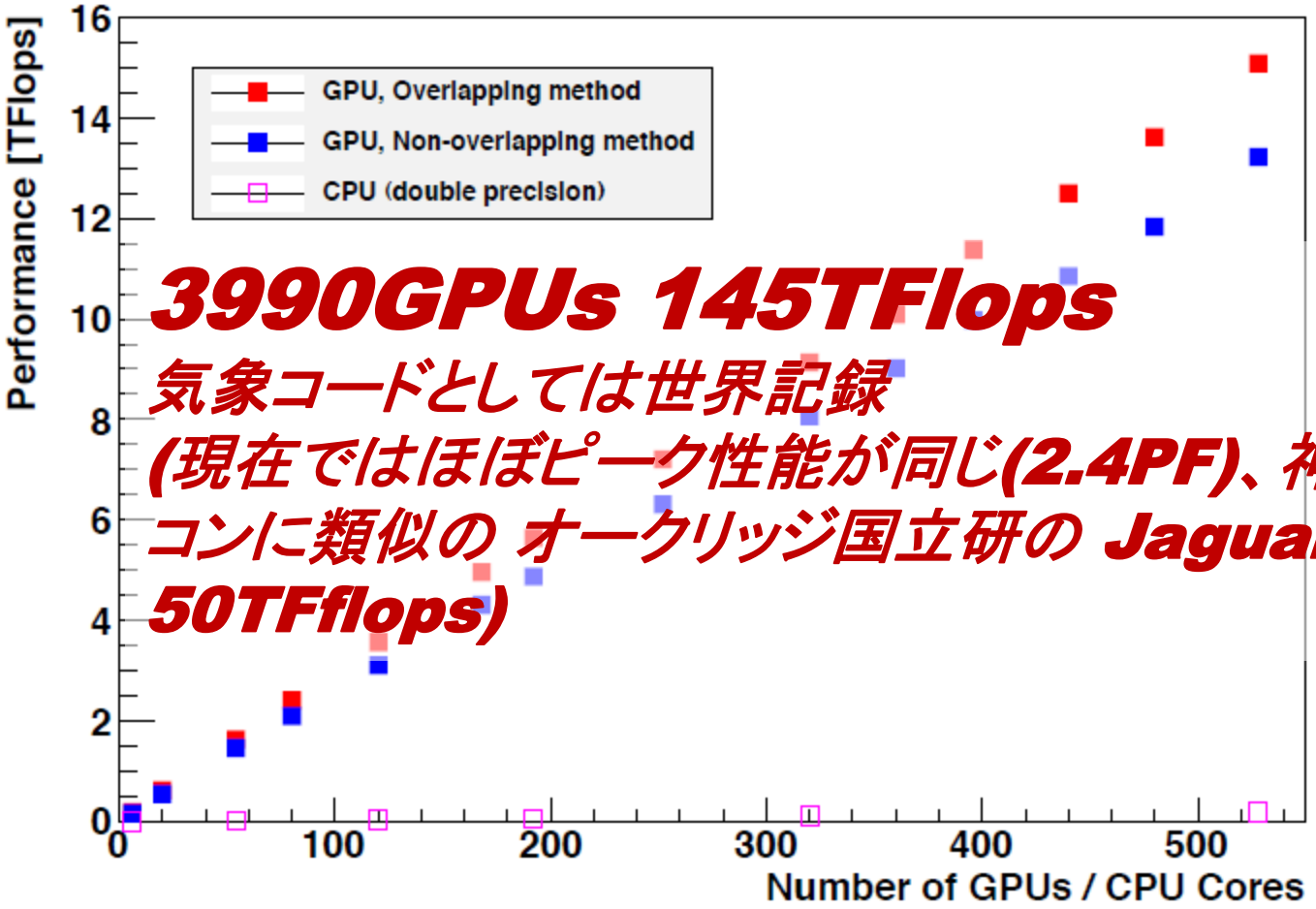




Typhoon



ASUCA Multi GPU Performance Supercomputing 2010で発表



3990GPUs 145TFlops

気象コードとしては世界記録

(現在ではほぼピーク性能が同じ(2.4PF)、神戸ペタコンに類似の オークリッジ国立研の Jaguar の 50TFflops)

対Jaguar電力性能比 8-12倍

コスト性能比10倍以上

MUPHY : Multi-Physics Simulator

[Massimo Bernaschi]

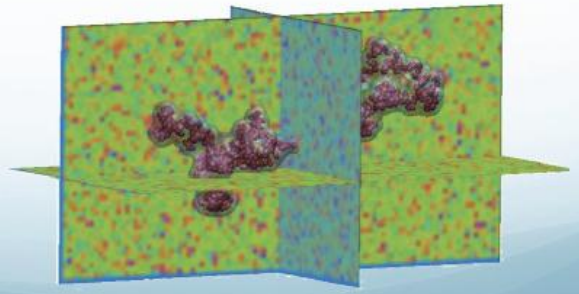
road spectrum of ***fluid-particle*** coupling mechanisms

brary of particle and molecular representations

brary of fluid types

o-polymers, colloidal suspensions, gels, biofluids, …

ulti-Platform

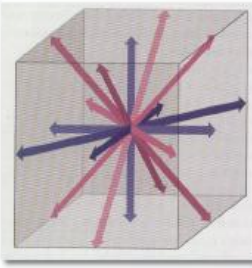


Blood Plasma via Lattice Boltzmann

From (minimal) Boltzmann equation

$$(\partial_t + \vec{v}\partial_x) f(\vec{x},\vec{v},t) = -\omega(f - f^{eq})(\vec{x},\vec{v},t)$$

Collision + Streaming of a set of discrete velocities



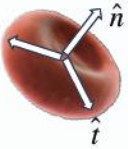
Superset of the Navier-Stokes dynamics

Exact streaming (no self-convected): uniform mesh

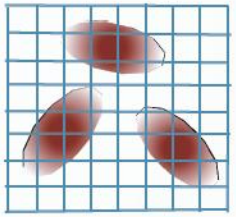
Complexity O(N)

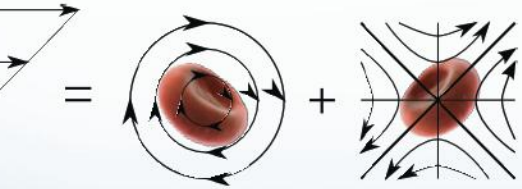
Enable complex geometries

Ellipsoidal Suspensions of RBC



$$\tilde{\delta}_{\xi_\parallel \xi_\perp}(x) = \prod_{\alpha=x,y,z} \tilde{\delta}_{\xi_\alpha}((Qx)_\alpha)$$





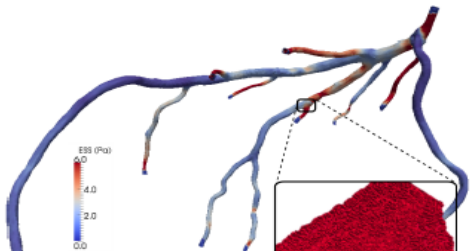
- ✓Tensorial drag
- ✓Tank Treading
- ✓Complex Rheology

Strip off DOFs: O(10) per RBC

MUPHY Performance on the TiTech GPU cluster

The performance of 1536 GPUs for the Lattice Boltzmann kernel is the same as the whole Jülich Bluegene/P system (294712 cores!)

# of GPUs	time	efficiency
256	76.16	N.A.
512	38.52	98.86%
1024	19.95	95.37%
1536	13.43	94.43%



x3 40 rack BG/P

# of GPUs	time
256	648.23
512	327.97

LB and Molecular Dynamics

呼吸器のシミュレーション

【東北大学医学部との共同研究】

Lattice Boltzmann Method

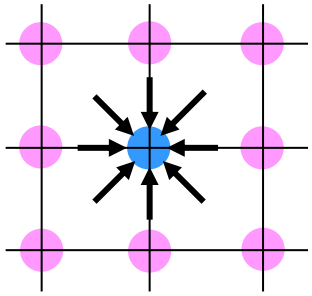
$$\frac{\partial f_i}{\partial t}+\mathbf{e}_i\cdot\nabla f_i=-\frac{1}{\lambda}\left(f_i-f_i^{eq}\right) \qquad f_i^{eq}=\rho w_i\left[1+\frac{3}{c^2}(\mathbf{e}_i\cdot\mathbf{u})+\frac{9}{2c^4}(\mathbf{e}_i\cdot\mathbf{u})^2-\frac{3}{2c^2}(\mathbf{u}\cdot\mathbf{u})\right]$$

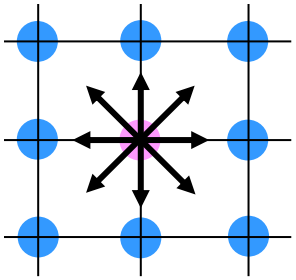
Collision step:

$$f_i(\boldsymbol{x}+\boldsymbol{e}_i\Delta t,t+\Delta t)=\bar{f}_i(\boldsymbol{x},t)$$

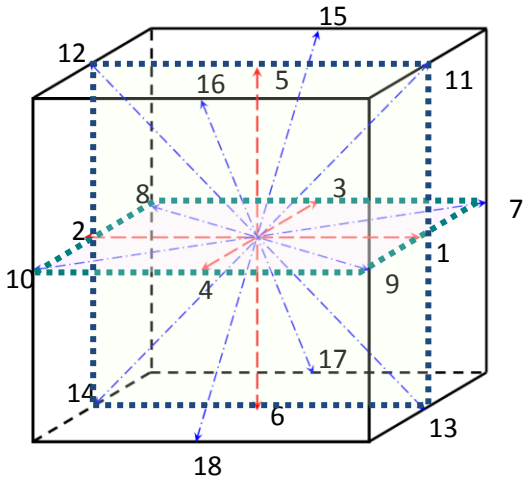
Streaming step:

$$\bar{f}_i(\boldsymbol{x},t)=f_i(\boldsymbol{x},t)-\frac{1}{\tau}\left(f_i(\boldsymbol{x},t)-f_i^{eq}(\boldsymbol{x},t)\right)$$

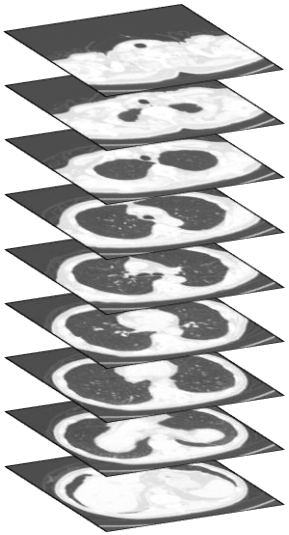




D3Q19

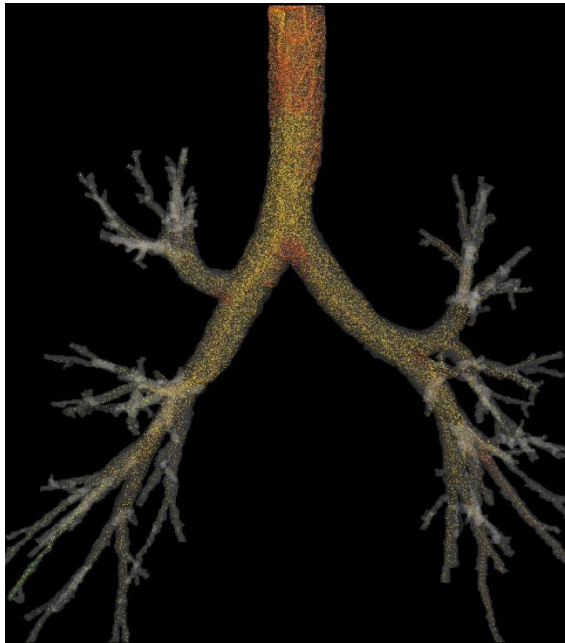


X-Ray CT images

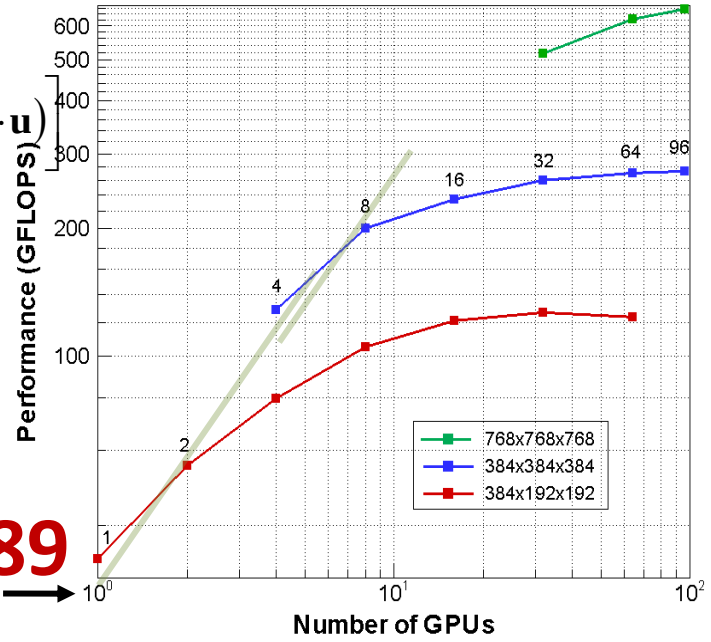


Airway structure Extraction

Lattice Boltzmann GPU computing



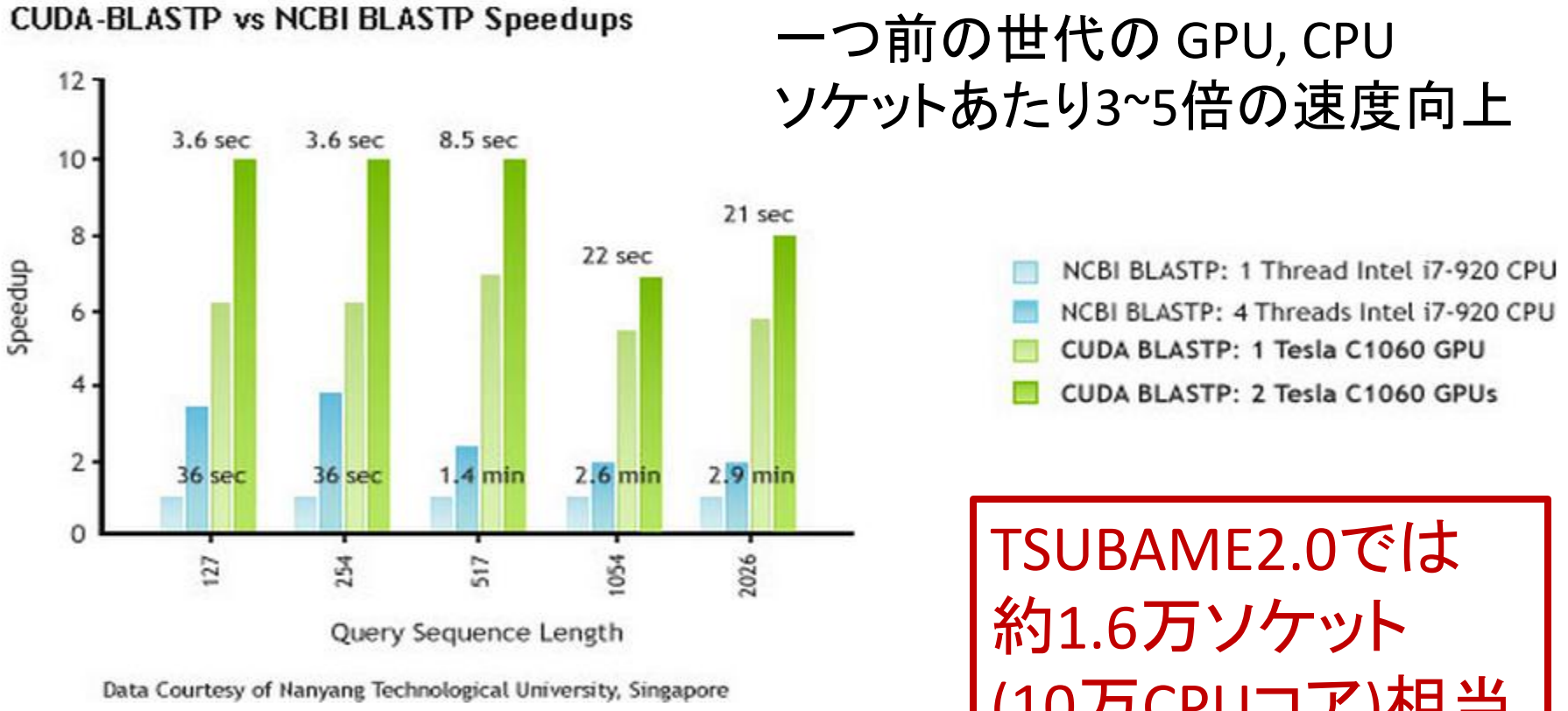
Collaboration with Tohoku University



× 89

バイオインフォ：遺伝子相同性検索BLAST on GPUs

- CUDA-BLASTP (NTU)
  http://www.nvidia.com/object/blastp_on_tesla.html



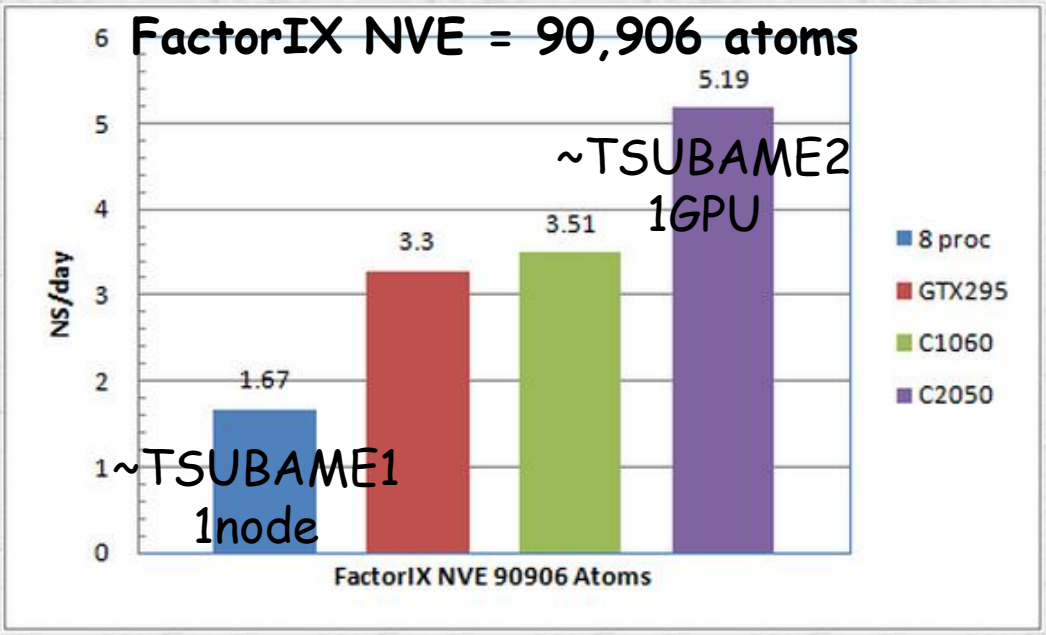
一つ前の世代の GPU, CPU
ソケットあたり3~5倍の速度向上

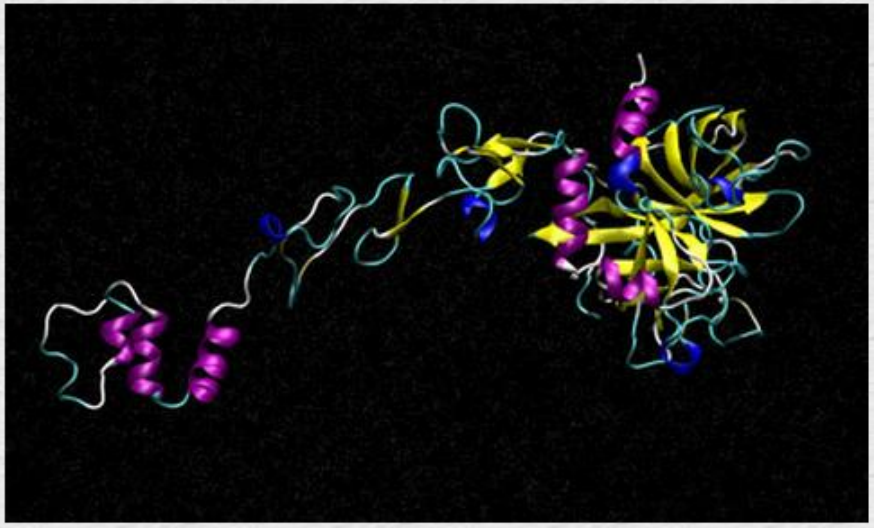
TSUBAME2.0では
約1.6万ソケット
(10万CPUコア)相当

- GPU-BLAST (CMU)
  http://eudoxus.cheme.cmu.edu/gpublast/gpublast.html
  “4 times speedup on Fermi GPUs”

分子動力学：GPU Amber 11

- ベンチマークでGPU1枚あたり8コアのCPUノードの4-20倍高速
  - マルチGPU版は「9月末に出る」はずだったがまだ出ていない。
- 精度は「問題ない」そうだが、今後検討の余地あり
- 平均10倍とすると、TSUBAME2.0では**30万コア**相当

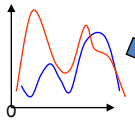




JST-CREST ULP-HPCの成果適用による スパコン・クラウド情報基盤のウルトラグリーン化

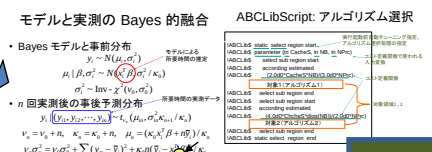
JST-CREST *Ultra Low Power HPC (2006-2013)*
スパコンの1000倍の性能電力向上を目指す基礎研究

ULP-HPC SIMD-Vector (GPGPU, etc.)

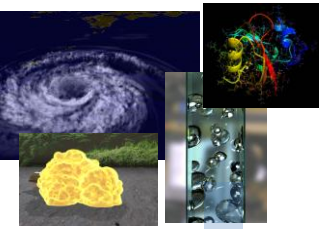


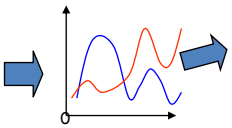
省電力高性能
ソフト機構・モデル化

(提案3) 自動チューニング共通基盤

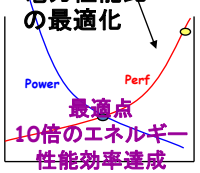


(提案1) ULP-HPC用次世代SW/HW要素利用
技術およびモデル化





省電力高性能
アルゴリズム・モデル化



2016年 TSUBAME
1/1000に

(提案2) 超省電力型のHPCアプリケーション及びアルゴリズム

(3) JST ULPHPC 基礎研究の適用によるスパコン用超省電力ミドルウェア・アプリ等

(2) TSUBAME2.0 (2010) 2.4PFlops
トップクラスの電力性能

Nvidia Maxwell
(2013)追加
TSUBAME2.5 ~10PF
1MWの維持

冷却用CO_2
排出の大幅削減

TSUBAME2.0/3.0

大気チャンバー
兼高効率電源設備

自然
大気吸気

フィルター

冷気

-TSUBAME3.0への成果-
2014-5年 TSUBAME3.0
数十PFlops ---TSUBAME1.0比で
数百倍の性能電力比

(1)自然大気冷却等 年平均PUE ~ 1.1

ULPHPCにおけるスパコン大規模高信頼性の研究

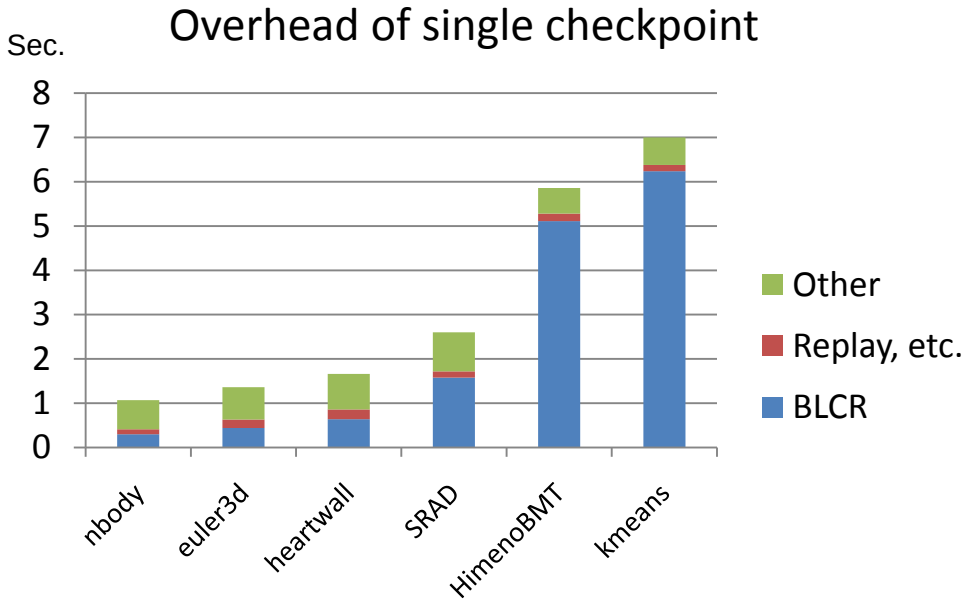
NVCR : A Transparent Checkpoint/Restart for CUDA

CUDA Checkpoint using BLCR (Takizawa, 2009)
- Save all data on CUDA resource.
- Destroy all CUDA context.
- BLCR
- Reallocate CUDA resource.
- Restore data.

After reallocation, the address or handle may differ from previous one.
Handles are Opaque pointers : can be virtualized.
Addresses must be visible for user application.

Our answer is `REPLAY'.
NVCR records calls of all APIs related to device memory address such as;
cuMem*(), cuArray*(), cuModule*().

We added some optimization of reducing the API records. For example, locally resolved alloc-free pairs are removed from the record.



The time for Replaying is quite negligible compared with the main BLCR procedure to write image to HDD.

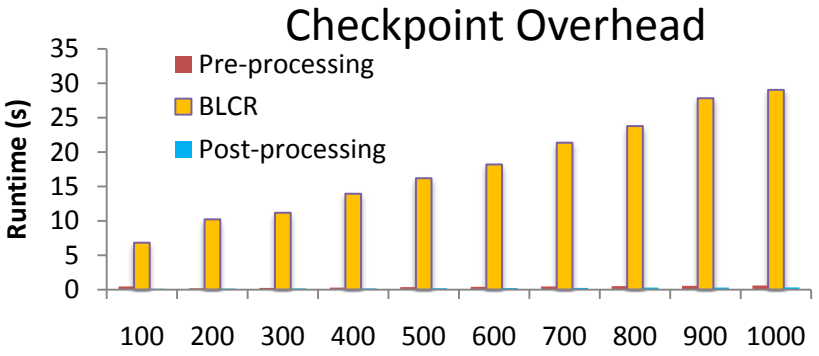
NVCR supports
- All CUDA 3.0 APIs
- OpenMPI
- Both CUDA runtime and CUDA driver API.

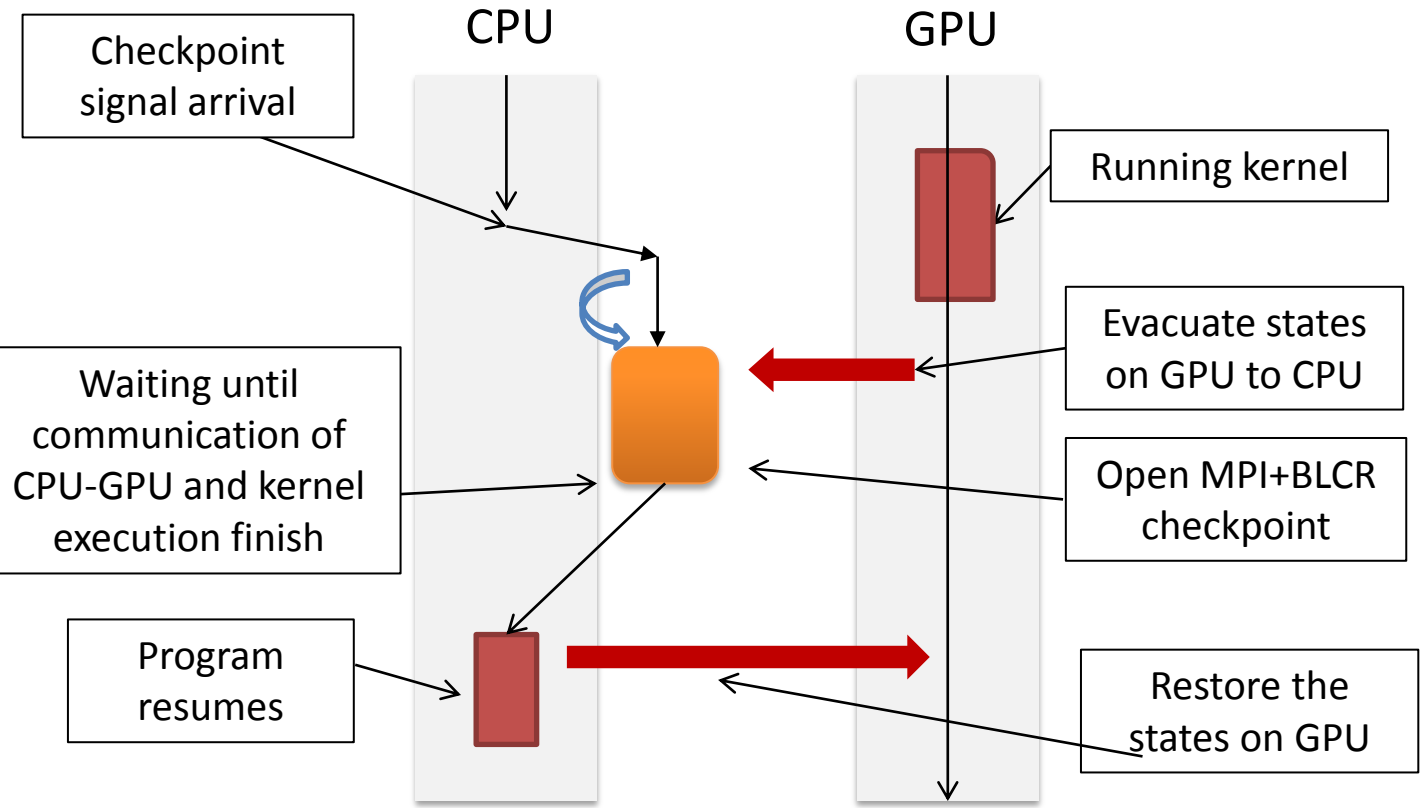
CUDA pinned memory is supported partially … Full support requires NVIDIA's official support.

MPI CUDA Checkpointing

Checkpointer for MPI+CUDA applications

- Very important in large-scale GPU systems
- Supports a majority of the CUDA RT API
- Based on BLCR and OpenMPI



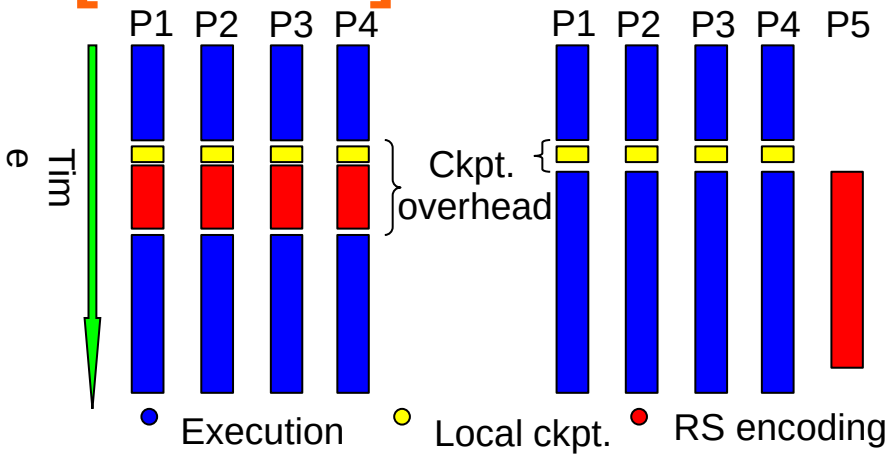


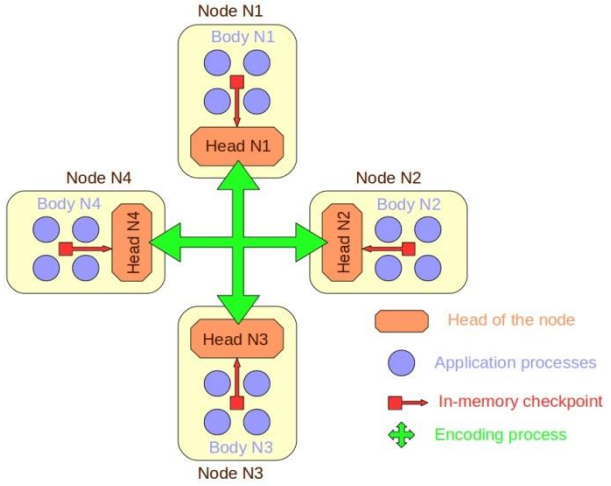
- GPU state saving and restoring has little overall performance impact.
- Can be significantly faster with our scalable checkpointing algorithms

More details at the Research Poster Session

Hybrid Diskless Checkpoint

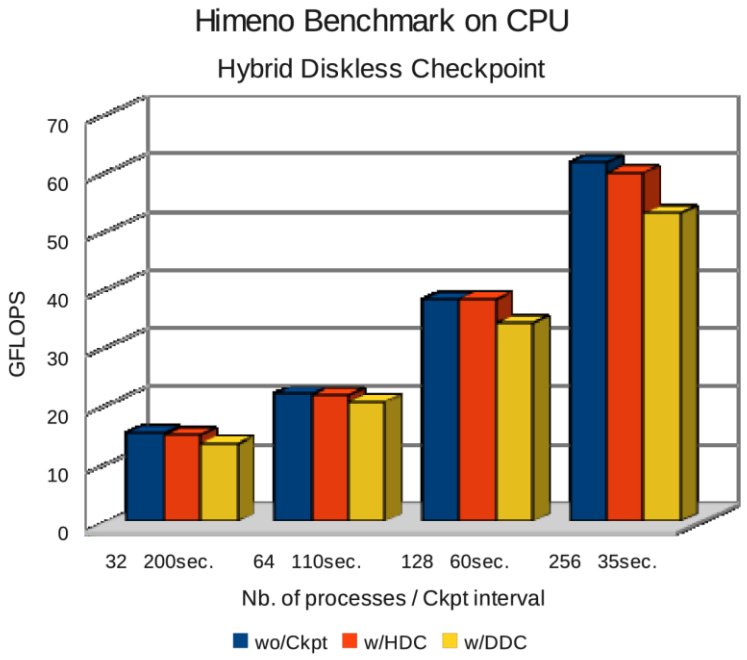
- ***Problem:*** *Decrease the ckpt. overhead with erasure codes on large-scale GPU systems.*
- Scalable encoding algorithm and efficient group mapping **[CCGrid10]**
- Fast encoding work on idle resource (CPU/GPU) **[HiPC10]**

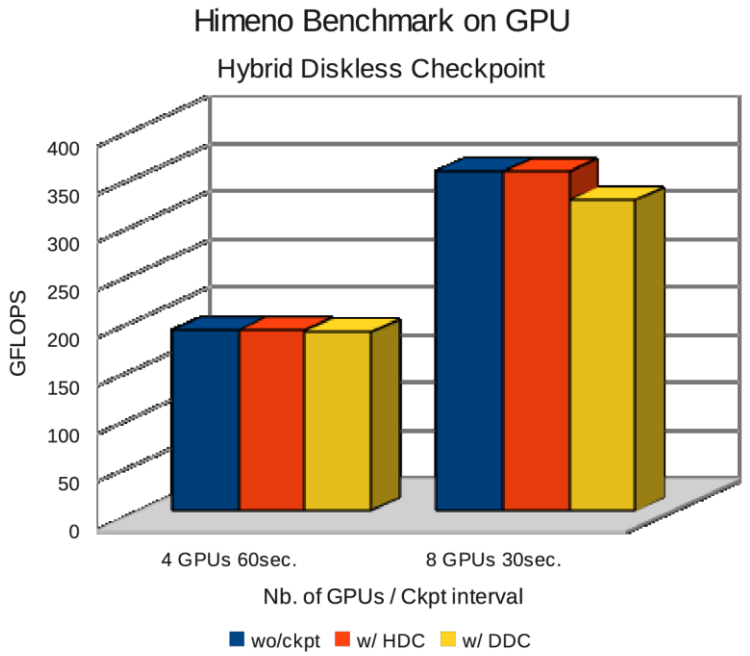




Hybrid Diskless Checkpoint

- Less than 3% of ckpt. overhead using idle resources





- We are currently combining this technique with our MPI CUDA Checkpoint.

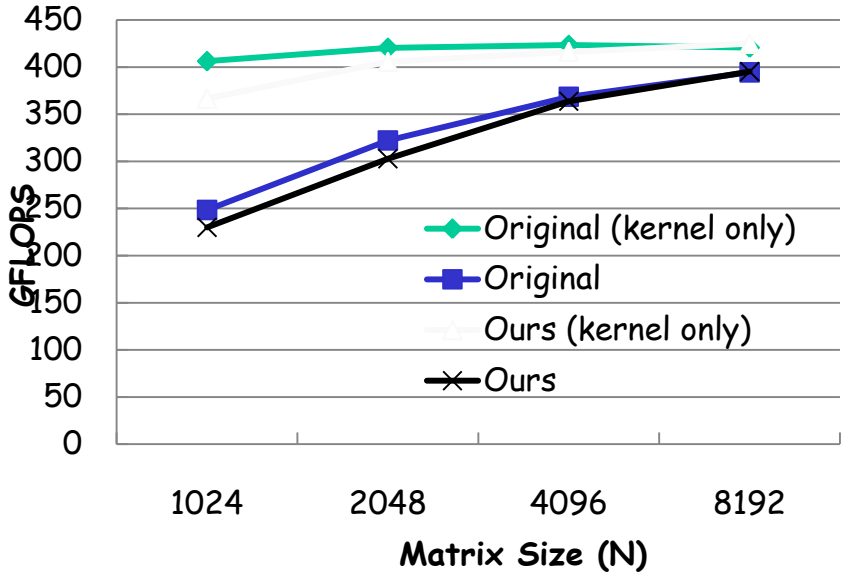
Software Framework for GPGPU Memory FT [IEEE IPDPS 2010]

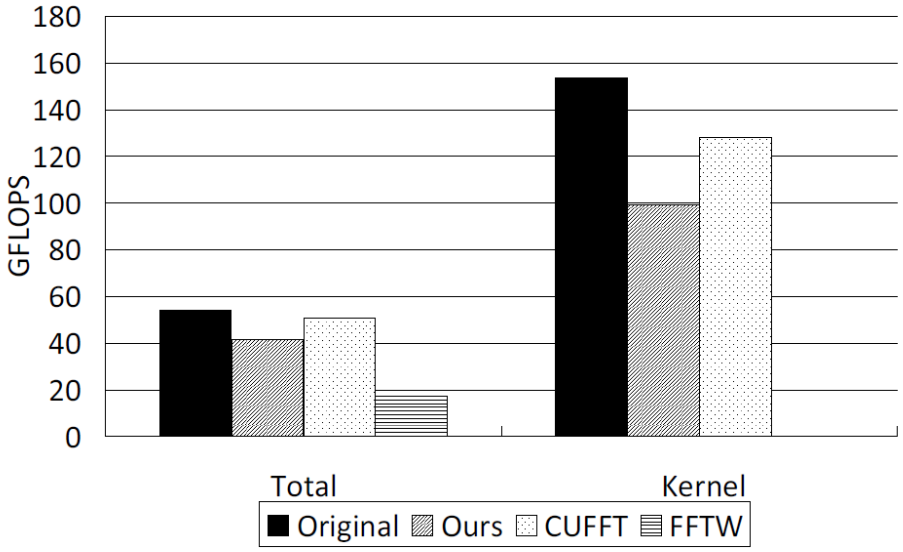
- Error detection in CUDA global memory + Checkpoint/Restart
- Works with existing NVIDIA CUDA GPUs

Lightweight Error Detection

- *Cross Parity* for 128B blocks of data
  - Detects a single-bit error in a 4B word
  - Detects a two-bit error in a 128B block
- No on-the-fly correction → Rollback upon error

Exploit the latency hiding capability of GPUs for data correctness





Exaflopsへ向けた我が国のeScience インフラ・基盤センターへの提言

- 世界レベルのLeadership Computing and Data Facilityとしての性能ロードマップ策定と、基盤センター群による競争・協調・多様性を伴った年次の着実な遂行：HPCIのあるべき型
  - 国家レベルでの年次スケーリングの目標設定
  - 多様性と競争原理の適用➔センター毎の定額予算からの脱却
- HPCIの資源センターが自らの計算科学・計算機科学・応用数学を連動するHPC研究開発・人材育成・産学連携体制
  - 理研NLPや研究所を中心とした連携・人材交流
  - 基礎研究=>運用実験=>(リーダーシップ)実運用
  - 「カタログからマシンを買うだけのセンター」はいらない
- 「ガラパゴス」から「国際連携」へ
  - LHCや国際宇宙ステーションに学べ

国際Exascale SW Project
http://www.exascale.org/
筑波 10/19-21/2009